

Accurate Prediction of Phase Transitions in Compressed Sensing via a Connection to Minimax Denoising

David Donoho*, Iain Johnstone* and Andrea Montanari*,†

November 7, 2011

Abstract

Compressed sensing posits that, within limits, one can undersample a sparse signal and yet reconstruct it accurately. Knowing the precise limits to such undersampling is important both for theory and practice.

We present a formula precisely delineating the allowable degree of undersampling of generalized sparse objects. The formula applies to Approximate Message Passing (AMP) algorithms for compressed sensing, which are here generalized to employ denoising operators besides the traditional scalar shrinkers (soft thresholding, positive soft thresholding and capping). This paper gives several examples including scalar shrinkers not derivable from convex optimization – the *firm* shrinkage nonlinearity and the *minimax* nonlinearity – and also nonscalar denoisers – block thresholding (both block soft and block James-Stein), monotone regression, and total variation minimization.

Let the variables $\varepsilon = k/N$ and $\delta = n/N$ denote the generalized sparsity and undersampling fractions for sampling the k -generalized-sparse N -vector x_0 according to $y = Ax_0$. Here A is an $n \times N$ measurement matrix whose entries are iid standard Gaussian. The formula states that the phase transition curve $\delta = \delta(\varepsilon)$ separating successful from unsuccessful reconstruction of x_0 by AMP is given by:

$$\delta = M(\varepsilon | \text{Denoiser}),$$

where $M(\varepsilon | \text{Denoiser})$ denotes the per-coordinate minimax mean squared error (MSE) of the specified, optimally-tuned denoiser in the *directly observed* problem $y = x + z$. In short, the phase transition of a noiseless *undersampling* problem is identical to the minimax MSE in a *denoising* problem. We derive this formula from State Evolution and we present numerical results validating the formula in a wide range of settings.

The above formula generates numerous insights, including: (a) in the scalar sparsity case, that ℓ_1 minimization is very close to optimal among all possible sparsity seeking algorithms, including those deriving from nonconvex optimization; (b) in the block sparsity case, block soft thresholding is dramatically outperformed by James-Stein block thresholding for large block-sizes; block James-Stein recovers block sparse signals for any $\delta > \varepsilon$, provided the blocklength is above a finite value $B_*(\delta, \varepsilon)$. (c) in the nonseparable cases of monotone regression and TV denoising, the sparsity-undersampling phase transition obtained by AMP tailored to the generalized sparsity is definitely better than the phase transition achievable using merely ℓ_1 minimization of coefficients in the Haar basis.

*Department of Statistics, Stanford University

†Department of Electrical Engineering, Stanford University

Key Words: Approximate Message Passing. Lasso. Group Lasso, Joint Sparsity, James-Stein, Minimax Risk over Nearly-Black Objects. Minimax Risk of Soft Thresholding. Minimax Risk of Firm Thresholding. Minimax Shrinkage. Nonconvex penalization. State Evolution. Total Variation Minimization. Monotone Regression.

Acknowledgements. NSF DMS-0505303, NSF DMS-0906812, NSF CAREER CCF-0743978

Contents

1	Introduction	4
1.1	AMP Framework	4
1.2	Minimax MSE for Denoising	5
1.3	Phase Transition for AMP	6
1.4	This Paper	7
1.5	Related Literature	9
2	Scalar-Separable Denoisers	10
2.1	Basic Simplifications	10
2.2	Minimax MSE of a Shrinker	10
2.3	Firm Shrinkage	11
2.4	Minimax Shrinkage	12
2.5	Empirical Phase Transition Behavior	15
3	Block-Separable Denoisers	16
3.1	Block Soft Thresholding	18
3.2	Block James-Stein	18
3.3	Empirical Phase Transition Behavior	22
4	Derivation of the Phase Transition by State Evolution	22
4.1	State Evolution	24
4.2	State Evolution Phase Transition	26
4.3	Non-Covergence of State Evolution	27
5	Nonseparable Denoisers	28
5.1	Monotone Regression	28
5.2	Total Variation Minimization	28
6	Discussion	29
6.1	Clarifications on AMP	29
6.2	Phase Transitions for other Algorithms	30
6.3	Interpretation as Penalized Optimization for Separable Denoiser	32
6.4	Comparison to Traditional (δ, ρ) Phase Diagrams	36
6.5	Contributions of This Paper	37

A	Classical Cases	38
B	Calculation of Minimax MSE	41
C	Calculation of Block Soft Thresholding Minimax MSE	44
C.1	Asymptotics for Large Blocksize	45
D	Non-Convergence of State Evolution	46
E	Minimax MSE for Monotone Regression	47
F	Minimax MSE for Total Variation Minimization	51
G	Computational Details	53
H	Finite-N Scaling	55
H.1	Offsets Decay Toward Zero	55
H.2	Transitions Sharpen	55
I	Risk Function Properties	56
I.1	Block Soft Thresholding	57
I.2	Positive-Part James-Stein Shrinkage	58
I.2.1	Risk at 0	58
I.2.2	Monotonicity of $R(\mu)$	59
I.2.3	Superquadraticity of $R(\mu)$	59

1 Introduction

In the noiseless compressed sensing problem, we are given a collection of linear measurements of an unknown vector x_0 :

$$y = Ax_0. \quad (1.1)$$

Here the measurement matrix A is n by N , $n < N$, and the N -vector x_0 is the object we wish to recover. Both y and A are known, while x_0 is unknown; we seek to recover an approximation to x_0 .

Since $n < N$, the equations are underdetermined; it seems hopeless to recover x_0 in general, but in compressed sensing one also assumes that the object is *sparse* in the appropriate sense. Suppose that the object is known to be k -sparse, i.e. to have k nonzero entries. If the problem size (k, n, N) is large, many recovery algorithms exhibit the phenomenon of *phase transition*.

Let $\varepsilon = k/N$ and $\delta = n/N$ denote the sparsity and undersampling parameters, respectively. Hence $(\varepsilon, \delta) \in [0, 1]^2$ defines a phase space for the different kinds of limiting situations we may encounter as (k, n, N) grow large. For a variety of algorithms and Gaussian matrices A with iid entries, one finds that this phase space can be partitioned into two phases: “success” and “failure”. Namely, for a given algorithm \mathcal{A} and given sparsity fraction ε , there exists a critical fraction $\delta(\varepsilon, \mathcal{A})$ such that if the sampling rate δ is larger than the critical value, $\delta > \delta(\varepsilon, \mathcal{A})$, then the algorithm is successful in recovering the underlying object x_0 with high probability¹, while if $\delta < \delta(\varepsilon, \mathcal{A})$ the algorithm is unsuccessful, also with high probability. Here, in the interesting cases, $\delta(\varepsilon, \mathcal{A}) < 1$, which means that it is indeed possible to undersample and still recover the underlying object; in fact $\delta(\varepsilon, \mathcal{A})$ shows precisely the limits of allowable undersampling. By now a large amount of empirical and theoretical knowledge has been compiled about the phase transitions exhibited by different algorithms; we refer the reader to [DT10c, Don06, DT09a, XH10, BCT11, Sto09, KWT09, DMM09, DMM11, Wai09]. In a parallel line of work, a number of sufficient conditions under which undersampling is possible using deterministic matrices have been studied, see e.g. [CT05, BRT09, BGI⁺08, Can06].

It is however fair to say that the research focused so far on ‘unstructured’ notions of sparsity whereby k simply counts the number on non-zero entries in x_0 . (We refer to Section 1.5 for an overview of related literature.) On the other hand, applications naturally lead to ‘structured’ notions of sparsity. This paper applies an algorithm framework - *Approximate Message Passing* (AMP) - to derive specific algorithms applicable to a variety of compressed sensing settings, including block and structured sparsity, convex and nonconvex penalization, and develops a single unifying formula that, specialized to each instance, gives the actual phase transition that we observe in practice. To give a preview of our results, we first recall some facts about AMP and statistical decision theory.

1.1 AMP Framework

Suppose \mathbf{X} is an unknown signal vector in \mathbf{R}^N that is of interest to us, and that we observe $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$, where $\mathbf{Z} \in \mathcal{N}(0, \sigma^2 I_N)$ is white Gaussian noise (here and below I_m denotes the $m \times m$ identity matrix). In other words, we have *direct* but noisy measurement of \mathbf{X} (in contrast to Eq. (1.1) above, where we had indirect incomplete but noiseless observations). Let $\eta(\cdot; \tau, \sigma) : \mathbf{R}^N \mapsto$

¹Throughout the paper, we will write that an event holds with high probability (w.h.p.) if its probability converges to 1 in the large system limit $N, n \rightarrow \infty$ with $\delta = n/N$ and $\varepsilon = k/N$ fixed.

\mathbf{R}^N be a *denoising* procedure, designed to recover objects \mathbf{X} typical in our application. In other words, if the observation takes value $\mathbf{Y} = y$, then we estimate the signal as $\eta(y; \tau, \sigma)$. The denoiser η depends parametrically on the noise level σ , and on τ that denotes one or more tuning parameters, to be optimized over.

A simple example is *soft thresholding*. Define the scalar nonlinearity $\eta_\lambda^{soft} : \mathbf{R} \mapsto \mathbf{R}$ where $\eta_\lambda^{soft}(y) = \text{sgn}(y)(|y| - \lambda)_+$. Then, for $y \in \mathbf{R}^N$, we define the denoising operator by applying the same nonlinearity componentwise, with $\lambda = \tau\sigma$: $\eta(y; \tau, \sigma) = ((\eta_{\tau\sigma}^{soft}(y_i))_{i=1}^N)$. This is appropriate to recover objects known to be sparse, and will be used as a running example in what follows.

Approximate Message Passing is an iterative scheme that takes a simple denoising procedure and turns it into a scheme for solving the compressed sensing problem. Working now in the setting of Eq. (1.1), it starts from $x^0 = 0$, and proceeds for iterations $t = 1, 2, \dots$ by maintaining a *current reconstruction* x^t and a *current working residual* z^t , and adjusting these iteratively. At iteration t , it forms a vector of *current pseudo-data* $y^t = x^t + A^* z^t$ and the next iteration's estimate is obtained by applying η to the current pseudo-data:

$$y^t = x^t + A^* z^t, \quad (1.2)$$

$$x^{t+1} = \eta(y^t; \tau, \sigma_t), \quad (1.3)$$

$$z^{t+1} = y - (Ax^{t+1}) + \Omega_t z^t. \quad (1.4)$$

Here σ_t is the current noise level, and Ω_t is a scalar. The computation of σ_t and Ω_t is described below.

Conceptually, AMP constructs an artificial denoising problem at each iteration and solves it using a standard denoising procedure. Thus it solves a compressed sensing problem by successive denoising.

For now, this description should be sufficient, save for two remarks:

1. The specifics of the construction are absolutely crucial for the results of this paper. These are embedded in the specification of the scale factors Ω_t and σ_t ; see Section 6.1.
2. The above algorithm framework was originally derived in [DMM09, DMM10] in the case of a *separable* denoiser η , i.e. a denoiser acting independently on each coordinate. In that paper the derivation was based upon constructing a proper belief propagation message passing algorithm, and then obtaining the above algorithm as a first-order approximation. Specific separable denoisers corresponded to different choices of the prior in belief propagation.

A central point of this paper is that the form of the algorithm (1.2), (1.3), (1.4) is really more general and can be used in settings outside the original derivation.

1.2 Minimax MSE for Denoising

Again in the denoising setting, we have a random vector $\mathbf{X} \in \mathbf{R}^N$ of interest and observe direct noisy observations $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$. Without loss of generality, we can assume $\sigma^2 = 1$, and denote the corresponding denoiser by $\eta(\cdot; \tau) \equiv \eta(\cdot; \tau, 1)$. Assume that our denoiser is parameterized by τ ranging over the set Θ . Also assume that the underlying distribution ν_N of the vector \mathbf{X} belongs the class \mathcal{F} of probability distributions. It is convenient to evaluate the procedure η through the following minimax problem

$$M(\mathcal{F}|\eta) = \frac{1}{N} \min_{\tau \in \Theta} \max_{\nu_N \in \mathcal{F}} \mathbb{E}_{\nu_N} \|\mathbf{X} - \eta(\mathbf{Y}; \tau)\|_2^2. \quad (1.5)$$

In words, we tune the denoiser optimally to control the (per-coordinate) mean-squared error for typical signals from even the most unfavorable choice within our class \mathcal{F} .²

In our running example, the denoiser is coordinatewise soft thresholding, and the class $\mathcal{F}_{N,\varepsilon}$ is the class of probability distributions supported on vectors $\mathbf{X} \in \mathbf{R}^N$ with at most $N\varepsilon$ nonzeros. Let $M(\mathcal{F}_{N,\varepsilon}|\eta)$ denote the corresponding minimax MSE. It turns out that $M(\mathcal{F}_{N,\varepsilon}|\eta)$ can be characterized in terms of a scalar estimation problem. Consider the collection $\mathcal{F}_{1,\varepsilon}$ of univariate probability distributions which place a fraction $1 - \varepsilon$ of their probability at zero and a fraction ε at another distribution. Let X denote a scalar random variable with distribution $\nu \in \mathcal{F}_{1,\varepsilon}$, and let $Y = X + Z$, where $Z \sim \mathcal{N}(0, 1)$. Define minimax soft thresholding MSE is:

$$M_1(\varepsilon|\text{Soft}) = \inf_{\lambda} \sup_{\nu \in \mathcal{F}_{1,\varepsilon}} \mathbb{E}_{\nu} \{ (X - \eta_{\lambda}^{\text{soft}}(Y))^2 \}.$$

One can show that $M(\mathcal{F}_{N,\varepsilon}|\eta) = M_1(\varepsilon|\text{Soft})$. The function $\varepsilon \mapsto M_1(\varepsilon|\text{Soft})$ has been characterized in [DJ94].

In several other simple examples $M(\mathcal{F}|\eta)$ can be nicely evaluated:

- The positive-constrained case, where $\mathbf{X} \geq 0$. In this case η is non-negative soft thresholding, and $\mathcal{F} = \mathcal{F}_{N,\varepsilon,+}$ is the class of probability distributions supported on vectors $\mathbf{X} \in \mathbf{R}^N$ that are nonnegative and $N\varepsilon$ -sparse.
- The box-constrained case where $\mathbf{X} \in [0, 1]^N$, and we use capping $\eta(x) = \min(1, \max(x, 0))$ coordinatewise. Here $\mathcal{F} = \mathcal{F}_{N,\varepsilon,\square}$ is the class of probability distributions supported on vectors $\mathbf{X} \in [0, 1]^N$ that are at the bounds 0 and 1 except in a fraction ε of coordinates.

In these cases $M(\mathcal{F}|\eta)$ can be reduced to the calculation of quantities $M_1(\varepsilon|\text{SoftPos})$, $M_1(\varepsilon|\text{Cap})$ defined in complete analogy to $M_1(\varepsilon|\text{Soft})$. In this paper we will give several other calculations of $M(\mathcal{F}|\eta)$, going considerably beyond these examples.

1.3 Phase Transition for AMP

Our principal result is the following:

Phase Transition Formula for AMP. First, consider the denoising problem for denoiser η , and denote by $M(\mathcal{F}_{\varepsilon}|\eta) = M(\varepsilon|\eta)$ the corresponding minimax MSE for classes of signals in $\mathcal{F}_{\varepsilon}$.

Next, consider the noiseless compressed sensing setting, with signal x_0 sampled from the class $\mathcal{F}_{\varepsilon}$, and reconstruction through AMP based on denoiser η . In the large-system limit $N, n \rightarrow \infty$ with $n/N = \delta$, AMP will correctly recover the underlying object for pairs (ε, δ) in the sparsity/undersampling phase diagram provided

$$\delta > M(\varepsilon|\eta). \tag{1.6}$$

Viceversa, for $\delta < M(\varepsilon|\eta)$, recovery outside this region is not guaranteed, in the sense that, for objects sampled from the least-favorable distribution for $M(\varepsilon|\eta)$, AMP will not correctly recover the signal.

²In some cases the per-coordinate mean-squared error is instead defined as a limit $\lim_{N \rightarrow \infty} N^{-1} \mathbb{E} \|\mathbf{X} - \eta(\mathbf{Y}; \tau)\|_2^2$ across a sequence of problems of increasing size.

This formula encompasses the following known results [DMM09]:

- AMP soft thresholding (with optimally chosen tuning parameter) correctly recovers for $\delta > M_1(\varepsilon|\text{Soft})$.
- AMP positive soft thresholding (with optimally chosen tuning parameter) correctly recovers for $\delta > M_1(\varepsilon|\text{SoftPos})$.
- AMP capping correctly recovers for $\delta > M_1(\varepsilon|\text{Cap})$.

Appendix A spells out how these existing results fall under the aegis of Eq. (1.6).

1.4 This Paper

Our aim in this paper is to show that formula (1.6) is correct in settings extending far beyond the three examples just mentioned. In brief, we lay out several denoising problems, and in each one verify the general formula. This requires in each case (a) calculating the minimax MSE for a problem of statistical decision theory; (b) implementing AMP for compressed sensing with the given optimal denoising family; and (c) verifying empirically that phase transition does indeed occur at the precise sparsity/undersampling tradeoff indicated by the formula.

In following sections, we consider the following different denoisers.

Firm Shrinkage. Here the denoiser η applies coordinatewise shrinkage using the so-called *firm* shrinkage nonlinearity η_λ^{firm} . This is a piecewise linear function of a scalar variable with two tuning parameters $\lambda = (\lambda_1, \lambda_2)$. It thresholds the argument y to zero for $|y| < \lambda_1$, i.e. $\eta_\lambda^{firm}(y) = 0$ in this case, and it behaves as the identity $\eta_\lambda^{firm}(y) = y$ for $|y| > \lambda_2$. Between the two regimes, it interpolates linearly. We show that minimax MSE function $M_1(\varepsilon|\text{Firm}) < M_1(\varepsilon|\text{Soft})$ strictly, and by verifying the general formula, we show that the phase transition curve for optimally-tuned AMP firm shrinkage is very slightly better than the phase transition for optimally tuned AMP soft shrinkage.

Minimax Shrinkage. Here the denoiser η applies coordinatewise shrinkage using a *minimax* shrinkage nonlinearity, hence implicitly we are optimizing the mean square error over $\eta \in \{\text{all scalar nonlinearities}\}$. We calculate the minimax MSE function $M_1(\varepsilon|\text{Minimax})$, show that $M_1(\varepsilon|\text{Minimax}) < M_1(\varepsilon|\text{Firm})$ strictly, and verify that the phase transition curve for AMP minimax shrinkage is slightly better than the phase transition for both AMP soft or firm shrinkage.

Block Thresholding. Here the denoiser η applies blockwise shrinkage using either block soft thresholding (for blocklength B , the B -variate nonlinearity obeys $\eta_{B,\lambda}(y) = y \cdot (1 - \|y\|_2/\lambda)_+$) or block James-Stein Shrinkage (we refer to Section 3 for an explicit definition). We will compute the minimax MSE function $M_B(\varepsilon|\text{BlockSoft})$, and bound the minimax MSE function $M_B(\varepsilon|\text{JamesStein})$. We will verify that the phase transition curve for optimally-tuned AMP block shrinkage follows the general formula.

Notice that, as demonstrated numerically in [DMM09], and proved in [BM11b] in the case of Gaussian sensing matrices, soft-thresholding AMP reconstruction coincides with LASSO reconstruction (in the large system limit). By the above results firm-shrinkage AMP and minimax AMP both

outperform LASSO reconstruction. Correspondingly, it can be argued that blocksoft thresholding AMP coincides asymptotically with group LASSO, and hence James-Stein AMP outperforms the latter.

In all of the above examples, the denoisers are coordinatewise or at least blockwise separable, so there is minimal “structure” to the sparsity. We next consider examples where the denoiser has more subtle structure. We find that formula (1.6) applies more generally, once we generalize the notion of sparsity.

Monotone Regression. Here η finds the least-squares projection of Y onto the cone of monotone increasing functions. As “sparse” sequences, we consider monotone signals with at most $N\varepsilon$ points of increase.

Total variation minimization. Here η minimizes the residual sum of squares penalized by τ times the total variation of the signal. As “sparse” sequences we consider signals with at most $N\varepsilon$ points of change.

In both cases, we find well-defined phase transitions, precisely at the location predicted by the general formula (1.6).

These results have further implications, because AMP-based results are known to connect to important phase transitions studied previously by other techniques.

- *Phase transitions in high-dimensional geometry.* Phase transitions in combinatorial geometry, such as the weak neighborliness phase transition for random projections of the Platonic solids [DT10a], have been shown to be identical to the phase transitions of AMP in certain “corresponding” problems [DMM09]. Hence, the general AMP phase transition formulas, which in the monotone and total variation cases refer to specific polytopes, generate fresh information, previously unknown in combinatorial geometry.
- *Phase transitions in convex optimization.* We show here that phase transitions of several convex optimization procedures in compressed sensing agree to high accuracy with formula (1.6). Indeed this agreement was proved rigorously in the case of standard ℓ_1 -regularized regression [BM11b]. As a result we have interesting new formulas for phase transitions in compressed sensing with *structured sparsity*.
- *Phase transitions in non-convex optimization.* We study here two AMP algorithms which are not equivalent to any convex optimization procedure, and show that (1.6) describes their phase transition curves, which are better than the phase transition curves of the more popular convex optimization procedures.

Finally, one may ask: why should (1.6) be true? We show here that it can be derived from the validity of state evolution for AMP algorithms in general. In other words, when state evolution accurately describes the behavior of AMP algorithms, formula (1.6) can be expected to apply. In the case of separable denoisers, the correctness of state evolution follows from the general result of [BM11a] (the latter requires the denoiser to be locally Lipschitz continuous, a condition that is easily checked for all the examples considered here). This result can be further generalized to block-separable denoisers [Ran11, BLM12]. The much broader validity of formula (1.6) we document in this paper suggests that state evolution is valid in a much broader setting than previously claimed.

1.5 Related Literature

Approximate message passing algorithms for compressed sensing reconstruction were introduced in [DMM09]. They were largely motivated by the connection with message passing algorithms in iterative decoding systems [RU08], and with mean field methods in statistical physics [MM09] (in particular the cavity method and TAP equations). We refer to [DMM11] for a discussion of these connections.

The original AMP framework [DMM09, DMM10] included iterations of the form defined in Eqs. (1.2), (1.3), (1.4) whereby the denoiser is separable and identical across coordinates, namely

$$\eta_t(v) = (\eta_t(v_1), \eta_t(v_2), \dots, \eta_t(v_N)). \quad (1.7)$$

While this covers the Firm and Minimax shrinkage rules studied in this paper, it did not include the various non-separable denoisers we discuss below, namely the block, monotone and total variation denoisers. Further, in [DMM09], the phase transition behavior was validated numerically only for Soft, SoftPos and Cap denoisers, that are in correspondence with well studied convex optimization problems. The extension to a noisy linear model $y = Ax_0 + w$, with $w \in \mathbf{R}^n$ a vector of iid random entries was carried out in [DMM11]. We also refer to [Mon12] for an overview of this work.

Several papers investigate generalizations of the original framework put forward in [DMM09]. The paper [BM11a] defines a general class of approximate message passing algorithms for which the state evolution was proved to be correct. This included in particular all separable Lipschitz-continuous denoisers.

Rangan [Ran11] introduces a class of generalized approximate message passing (G-AMP) algorithms that cope with –roughly– two generalizations of the basic noisy linear model. First, the noisy measurement vector y can be a non-linear (random) function of the noiseless measurement Ax_0 . Second, each of the ‘coordinates’ of x_0 can itself be a –low dimensional– vector. Interesting applications of this framework were developed in [KGR11, Sch11]. Let us notice that G-AMP does not cover any of the non-separable cases treated here (even the block sparse example), and hence provides a generalization in an ‘orthogonal’ direction.

In a parallel line of work, Schniter applied AMP to a number of examples in which the signal x_0 has a structured prior [Sch10, SSS10, SPS10]. Inference with respect to the prior is carried out using belief propagation, and this is combined with AMP to compute *a posteriori* estimates. This type of application fits within the class of problems studied here, by choosing the denoiser η_t in Eq. (1.3) be given by the appropriate conditional expectation with respect to the signal prior. On the other hand, the general scheme provided by Eqs. (1.2), (1.3), (1.4) encompasses cases in which the denoiser is not derived from a specific prior.

Maleki et al. [MAYB11] used methods analogous to the ones developed here to study phase transitions for compressed sensing with complex vectors. This is closely related to the block-separable setting considered in Section 3 (there is however some difference in the structure of the sensing matrix).

Structured sparsity models are studied from a different point of view in [BCDH10, CHDB08, CICB10]. Those works focus on deriving sparsity models that capture a variety applications, and of convex relaxations that promote the relevant sparsity patterns. Reconstruction guarantees are proved under suitable ‘isometry’ assumptions on the sensing matrix.

Closer to our approach is a recent series of papers [CRPW11, RRN11, RRN11], considering general classes of structured signals under random measurements. Let us emphasize two important

differences with respect to our work. First, these papers only deal with convex reconstruction methods, while we shall analyze several nonconvex approaches and demonstrate improvements. Second, they derive reconstruction guarantees using concentration-of-measure arguments, while we derive exact asymptotics (essentially based on weak convergence), which enables us to unveil the key relation (1.6) between denoising and the compressed sensing phase transition.

2 Scalar-Separable Denoisers

In this section we study scalar-separable denoisers, i.e. denoisers that take the form

$$\eta_t(v) = (\eta_t(v_1), \eta_t(v_2), \dots, \eta_t(v_N)), \quad (2.1)$$

where we assume that the denoiser is symmetric across coordinates and, with a slight abuse of notation, we used the same symbol for the overall vector denoiser and for the scalar denoiser $\eta_t : \mathbf{R} \mapsto \mathbf{R}$. Often (but not always) η_t obeys $\|\eta_t(v)\|_2 \leq \|v\|_2$, so one commonly encounters the term ‘shrinker’ rather than ‘denoiser’.

2.1 Basic Simplifications

All of our scalar denoisers $\eta_t : \mathbf{R} \rightarrow \mathbf{R}$ take the form $\eta_t(Y) = \eta(Y; \tau, \sigma_t)$. Here $\tau \in \Theta$ denotes one or more tuning parameters that depend on the signal class but not on the iteration. Also σ_t is the effective noise level in observation Y . Under the state evolution formalism described in Section 4, at iteration t , in coordinate i , we have $Y = y_{t,i} \approx x_{0,i} + \sigma_t Z_i$ with $Z_i \sim \mathbf{N}(0, 1)$.

In fact, the scalar denoiser will be specified only for $\sigma_t = 1$, and extended through the scaling relation

$$\eta(y; \tau, \sigma_t) = \sigma_t \eta\left(\frac{y}{\sigma_t}; \tau, 1\right). \quad (2.2)$$

Notice that this choice is justified by the fact that the signal classes under consideration are invariant under rescaling. We will use the simpler notation $\eta(y; \tau) = \eta(y; \tau, 1)$ for the denoiser at noise level one. This section develops the machinery needed to apply formula (1.6) to the case of scalar denoiser operators.

2.2 Minimax MSE of a Shrinker

Consider the scalar observation $Y = X + Z$ where X is the unobserved quantity of interest, $Z \sim \mathbf{N}(0, 1)$ independently of X , and Y is observed. Denote by ν the distribution of X ; the mean-squared error (MSE) of the shrinker is defined so:

$$\text{MSE}(\eta(\cdot; \tau); \nu) = \mathbb{E}\left\{[\eta(X + Z; \tau) - X]^2\right\}.$$

Again, the random variables $Z \sim \mathbf{N}(0, 1)$ and $X \sim \nu$ are independent.

Obviously, the performance depends on both the tuning parameter τ and the distribution ν of X . The random variables X of interest are only occasionally nonzero. Let $\mathcal{F}_{1,\varepsilon}$ denote the collection 1-dimensional marginal distributions placing mass $(1 - \varepsilon)$ at the origin:

$$\mathcal{F}_{1,\varepsilon} = \{\nu : \nu \text{ is probability measure on } \mathbf{R} \text{ with } \nu(\{0\}) \geq 1 - \varepsilon\}.$$

We evaluate the performance of tuning constant τ by the worst-case performance it yields over $\mathcal{F}_{1,\varepsilon}$, namely $\max_{\mathcal{F}_{1,\varepsilon}} \text{MSE}(\eta(\cdot; \tau), \nu)$. The minimax tuning parameter $\tau^*(\varepsilon)$ solves the problem:

$$M(\varepsilon|\eta) = M(\mathcal{F}_{1,\varepsilon}|\eta) = \inf_{\tau \in \Theta} \sup_{\nu \in \mathcal{F}_{1,\varepsilon}} \text{MSE}(\eta(\cdot; \tau), \nu). \quad (2.3)$$

The minimax shrinkage procedure (within the class of $\eta(\cdot; \tau)$) is denoted $\eta^*(\cdot) = \eta(\cdot; \tau^*)$.

This general framework has been considered before in the context of soft thresholding $\eta(y; \tau) = \text{sgn}(y)(|y| - \tau)_+$, positive soft thresholding $\eta(y; \tau) = \eta^{\text{softpos}}(y; \tau) = (y - \tau)_+$, and also hard thresholding $\eta(y; \tau) = y1_{\{|y| > \tau\}}$, and the minimax MSE and minimax thresholds were derived and analyzed in [DDS92, DJ94]. Plots of the minimax soft threshold $\tau^*(\varepsilon)$ and the minimax MSE are available in [DMM09, DMM11]. Such plots also appear later in this paper as baselines for comparison of interesting new families, namely Firm Shrinkage and Minimax Shrinkage.

2.3 Firm Shrinkage

A frequently-voiced criticism of ℓ_1 minimization and soft thresholding is the tendency to shrink large values by more than warranted. In the mid 1990's, *firm* shrinkage was introduced to correct this tendency by Gao and Bruce [GB97]; the name reminds us that it is intermediate between soft and hard thresholding; continuous like soft, but not shrinking large values – just like hard. The firm nonlinearity $\eta_{\lambda}^{\text{firm}}(\cdot; \tau_1, \tau_2)$ (denoted above by $\eta_{\lambda}^{\text{firm}}$) has two threshold parameters $\tau = (\tau_1, \tau_2)$, whereby $0 < \tau_1 < \tau_2$. Below τ_1 the output is zero. Above τ_2 the output is the same as the input. Between these two thresholds, the output is determined by linear interpolation of the values at the thresholds. Formally:

$$\eta^{\text{firm}}(y; \tau_1, \tau_2) = \begin{cases} 0 & |y| < \tau_1, \\ \text{sgn}(y)(|y| - \tau_1)\tau_2/(\tau_2 - \tau_1) & \tau_1 < |y| < \tau_2, \\ y & |y| > \tau_2. \end{cases}$$

The family $\eta^{\text{firm}}(\cdot; \tau_1, \tau_2)$ contains soft and hard thresholding as limiting cases:

$$\eta^{\text{soft}}(y; \tau_1) = \lim_{\tau_2 \rightarrow \infty} \eta^{\text{firm}}(y; \tau_1, \tau_2), \quad \eta^{\text{hard}}(y; \tau_1) = \lim_{\tau_2 \rightarrow \tau_1} \eta^{\text{firm}}(y; \tau_1, \tau_2). \quad (2.4)$$

In this setting the general prescription for minimax MSE reduces to:

$$M(\varepsilon|\text{Firm}) = M(\mathcal{F}_{1,\varepsilon}|\eta^{\text{firm}}) = \inf_{0 < \tau_1 \leq \tau_2} \sup_{\nu \in \mathcal{F}_{1,\varepsilon}} \text{MSE}(\eta^{\text{firm}}(\cdot; \tau_1, \tau_2), \nu). \quad (2.5)$$

Implicit in this calculation is determination of the optimal thresholds $\tau_1^*(\varepsilon)$ and $\tau_2^*(\varepsilon)$. One can show that $0 \leq M(\varepsilon|\text{Firm}) \leq 1$, that M is monotone increasing with ε , that $M(\varepsilon) \rightarrow 1$ as $\varepsilon \rightarrow 1$ and that $M(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$.

Figure 1 shows the minimax MSE for firm shrinkage. The figure also shows similar results for soft and hard thresholding, for comparative purposes. Over the range presented, the minimax MSE for firm thresholding is *strictly smaller* than the MSE for hard or soft thresholding. Namely, over this range of ε ,

$$M(\varepsilon|\text{Firm}) < M(\varepsilon|\text{Soft}) < M(\varepsilon|\text{Hard}).$$

This in a sense validates certain traditional criticisms of soft thresholding, which is often said to shrink large values “too heavily”³.

³Note, however, that the use of hard thresholding instead of soft thresholding makes things worse!

ε	$M(\varepsilon \text{Hard})$	$M(\varepsilon \text{Soft})$	$M(\varepsilon \text{Firm})$	$M(\varepsilon \text{Minimax})$
0.010	0.0729	0.0612	0.0552	0.0533
0.025	0.1547	0.1231	0.1137	0.1093
0.050	0.2676	0.2039	0.1921	0.1841
0.100	0.4497	0.3288	0.3165	0.3025
0.150	0.5960	0.4279	0.4171	0.3983
0.200	0.7161	0.5111	0.5024	0.4802
0.250	0.8141	0.5829	0.5763	0.5516

Table 1: Minimax MSE of various scalar nonlinearities, across sparsity levels

Figure 2.3 shows the minimax thresholds. At least for $\varepsilon < 1/3$ we see clearly that $\tau_1^*(\varepsilon) < \tau_2^*(\varepsilon) < \infty$, so firm thresholding is preferred over the limiting cases of hard and soft thresholding.⁴ Figure 2.3 shows the corresponding minimax denoisers for specific values of ε .

2.4 Minimax Shrinkage

The previous example showed that a parametric family of shrinkers can improve on soft thresholding, and hence improve the predicted phase transition according to (1.6). The ultimate improvement one could make in this direction is to use the *globally* minimax nonlinear shrinker. This is the nonlinearity η that is minimax not within some parametric family, such as the soft thresholding or the broader firm thresholding family, but minimax over *all* measurable nonlinearities $\eta : \mathbf{R} \rightarrow \mathbf{R}$. Formally, let $\mathcal{L}(\mathbf{R})$ be the set of all measurable functions $\tau : \mathbf{R} \rightarrow \mathbf{R}$, for such a τ , set $\eta^{all}(x; \tau) \equiv \tau(x)$. The minimax MSE over this class is

$$\begin{aligned}
M(\varepsilon|\eta^{all}) &= \inf_{\tau \in \mathcal{L}(\mathbf{R})} \sup_{\nu \in \mathcal{F}_{1,\varepsilon}} \text{MSE}(\eta^{all}(\cdot; \tau), \nu) \\
&= \inf_{\eta \in \mathcal{L}(\mathbf{R})} \sup_{\nu \in \mathcal{F}_{1,\varepsilon}} \text{MSE}(\eta, \nu) \\
&\equiv M(\varepsilon|\text{Minimax}).
\end{aligned} \tag{2.6}$$

The calculation of this quantity uses a variety of ideas from minimax decision theory, for example in [Bic81, CS81, BC83, DDS92, Joh94b, Joh94a, DJ94]; details are given in Appendix B. A key point of this derivation is the characterization of the minimax nonlinearity as the MMSE Bayes rule (that is, the conditional expectation) for the so-called least-favorable prior. The least-favorable prior is the solution of Mallows' classical Fisher information problem [Mal78], solved here numerically.

Table 1, Figure 1 present numerical values associated with the solution of the minimax problem. Inevitably, $M(\varepsilon|\text{Minimax}) < M(\varepsilon|\text{Firm}) \leq M(\varepsilon|\text{Soft})$, i.e. optimizing over all nonlinearities yields a smaller mean square error than soft or firm thresholding. On the other hand, Table 1 shows that the improvements are typically of size 0.01 or smaller over the range $\varepsilon \in (0.01, 0.25)$. For very small ε it was pointed out in [DMM09] that [DJ94] showed

$$\lim_{\varepsilon \rightarrow 0} \frac{M(\varepsilon|\text{Minimax})}{M(\varepsilon|\text{Soft})} = 1.$$

⁴These are numerical results. Open question: for ε larger than the range studied here, does the minimax firm threshold have parameter $\tau_2(\varepsilon) = \infty$ reducing it to soft thresholding.

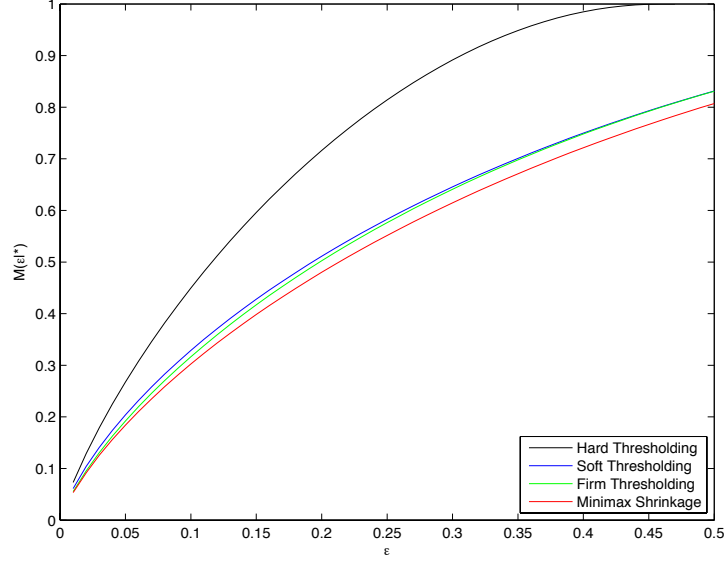


Figure 1: Minimax MSE of various scalar nonlinearities, as a function of sparsity parameter ε . Minimax MSE of firm thresholding is very close to that of soft thresholding for $\varepsilon > 1/3$.

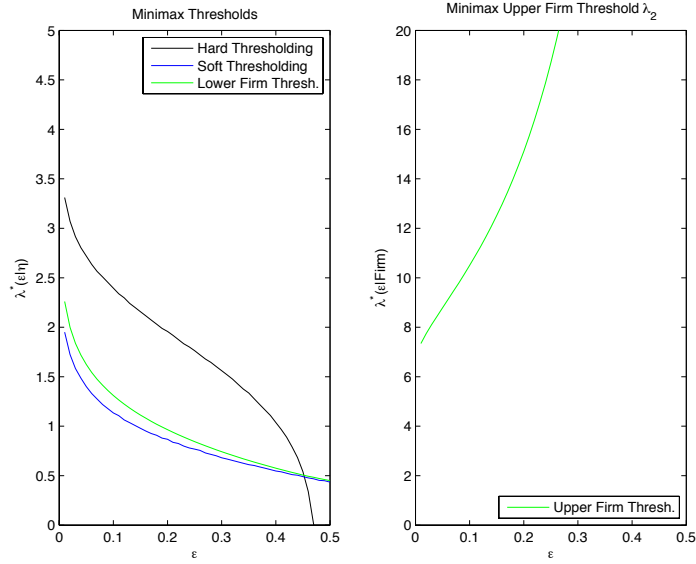


Figure 2: Minimax thresholds of various thresholding nonlinearities, as a function of sparsity ε . The sudden drop in the value of the hard threshold to zero near 0.45 coincides with the value of the minimax MSE in Figure 1 reaching 1.0. The numerical approximation to the minimax lower firm threshold approaches the minimax soft threshold as ε increases beyond 0.3, and the upper firm threshold increases rapidly.

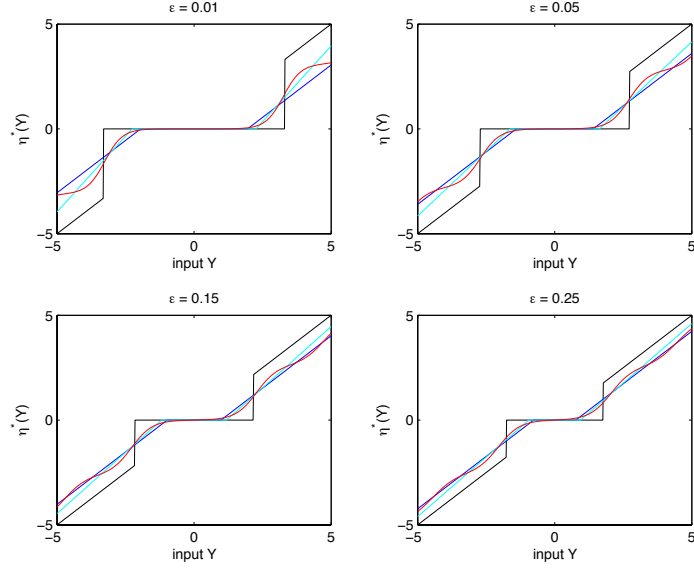


Figure 3: Minimax nonlinearities of various types, at sparsity $\varepsilon \in \{0.01, 0.05, 0.15, 0.25\}$. Black: hard thresholding; Blue: soft; Aqua: firm; Red: minimax shrinkage.

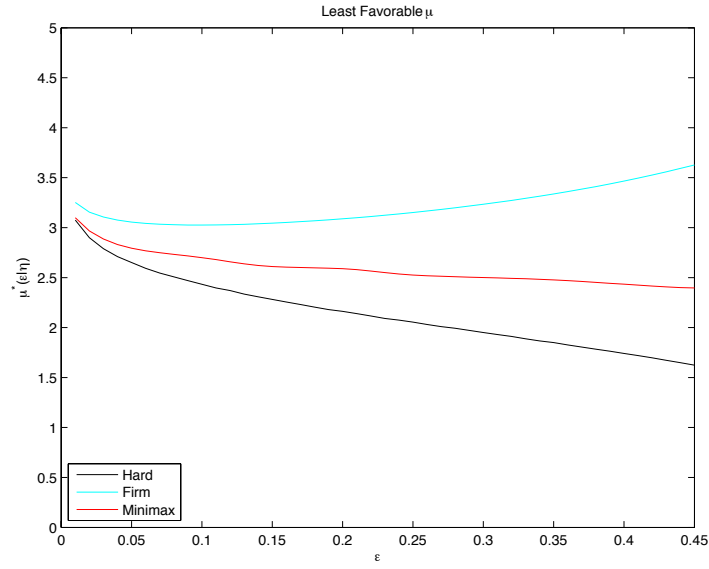


Figure 4: Least Favorable μ for various thresholding nonlinearities, as a function of sparsity ε . Note: soft thresholding has $\mu^*(\varepsilon|\text{Soft}) = \pm\infty$, not shown.

In the limit of extreme sparsity, there is nothing to be gained by completely general nonlinearities over soft thresholding! The improvement is non-vanishing, but moderate for ε non-vanishing.

2.5 Empirical Phase Transition Behavior

The research hypothesis driving this paper is that Eq. (1.6) describes the phase transition of AMP algorithms. In order to be completely explicit, we need to check the following predictions, for each nonlinearity η of interest

1. There exists a curve $\varepsilon \mapsto \delta(\varepsilon|\eta)$ such that for $\delta > \delta(\varepsilon|\eta)$ the corresponding AMP algorithm will typically succeed in reconstructing the unknown signal x_0 , and for $\delta < \delta(\varepsilon|\eta)$ the algorithm will typically fail.
2. The curve is related to the corresponding scalar denoising problem by $\delta(\varepsilon|\eta) = M(\varepsilon|\eta)$.

We now test this hypothesis for the firm and globally minimax nonlinearities $\eta \in \{\eta^{firm}, \eta^{all}\}$.

Our experiment was conducted along the same lines as [MD10, DT09b, DMM09, DMM11, BT10]. We considered a range of problem sizes $N \in \{1000, 2000, 4000\}$ and a range of sparsity parameters $\varepsilon \in \{0.01, 0.02, 0.05, .10, 0.15, 0.20, 0.25\}$. We considered a grid of δ values surrounding the predicted phase transition $\delta(\varepsilon|\eta)$, and ran $N_{\text{sample}} = 1000$ Monte Carlo reconstructions at each parameter combination. We declared “success” when the relative mean-squared error was below 1%:

$$\frac{\|\hat{x}^t - x_0\|_2^2}{\|x_0\|_2^2} < 0.01.$$

We used $t = 300$ iterations of AMP⁵.

This experiment generated data set for each optimal nonlinearity, containing, for each algorithm and each fixed ε , a list of values δ_i and empirical success fractions \hat{p}_i . The success fractions observed at $\delta > M(\varepsilon|\eta)$ were indeed typically better than 50% and at $\delta < M(\varepsilon|\eta)$ were typically worse than 50%.

To quantify this tendency, we fit a logistic regression

$$\text{logit}(\hat{p}_i) = \alpha + \beta(\delta_i - \delta(\varepsilon|\eta)), \quad (2.7)$$

where $\delta(\varepsilon|\eta) = M(\varepsilon|\eta)$ was computed analytically using ideas mentioned earlier. For each set of data corresponding to given (N, ε) and each non-linearity, we estimate α and β from the logit fit, leading to values $\hat{\alpha}, \hat{\beta}$. Using these quantities, we estimate the phase transition location as the value at which the probability \hat{p} of success is 50%. Using Eq. (2.7) this corresponds to $\hat{\alpha} + \hat{\beta}(\delta - \delta(\varepsilon|\eta)) = 0$, i.e. $\delta = \delta(\varepsilon|\eta) - (\hat{\alpha}/\hat{\beta})$. We are therefore led to define the offset between the empirical phase transition and the prediction $\delta(\varepsilon|\eta) = M(\varepsilon|\eta)$ as

$$\widehat{\text{PT}} = \widehat{\text{PT}}(N, \varepsilon) \equiv -\frac{\hat{\alpha}}{\hat{\beta}}. \quad (2.8)$$

In order to check the general relation provided by Eq. (1.6) we need to show that $\widehat{\text{PT}}(N, \varepsilon)$ tends to zero as N gets large, to within the statistical uncertainty. In Table 2 we report our results on the empirical phase transition, confirming that indeed the offset is small and decreasing with N .

Notes:

⁵In most cases the mentioned convergence criterion is reached after a much smaller number of iterations (roughly 20)

- We also calculated formal 95% confidence intervals for $\widehat{\text{PT}}$, indicating the tight control we have of the correct value.
- As in earlier studies [DT09b], we don't expect the empirical PT to lie exactly at the prediction, but instead to obey a relation of the form

$$\text{empirical PT}_N - \text{predicted PT}_{N \rightarrow \infty} \approx \frac{\text{const}}{N^\gamma} + \text{sampling error},$$

for some $\gamma \in (0, 1]$. Our data supports this relationship, with $\gamma \approx 1/3$. See Appendix H.

- Denoting by $\widehat{\beta}_N$ the fitted slope coefficient at dimension N , evidence that $\widehat{\beta}_N$ is increasing with larger N indicates that a sharpening of the phase transition is indeed occurring. Appendix H shows that $\widehat{\beta}_N \sim \sqrt{N}$ is consistent with our data.

3 Block-Separable Denoisers

We now turn to the case of block-structured sparsity, first introducing some notational conventions. We partition the vector $x = (x_1, x_2, \dots, x_N)$ into M blocks each of size B . Denoting by $\text{block}_m(x) = (x_{(m-1)B+1}, \dots, x_{mB})$ the m -th block, we write

$$x = (\text{block}_1(x), \dots, \text{block}_M(x)),$$

with, of course, $N = MB$.

A block-separable denoiser is a mapping $\eta_t : \mathbf{R}^N \rightarrow \mathbf{R}^N$ that decomposes according to the above partition:

$$\eta_t(x) = (\eta_t(\text{block}_1(x)), \dots, \eta_t(\text{block}_M(x))), \quad (3.1)$$

where, with an abuse of notation, we use the same symbol to denote the single-block denoiser $\eta_t : \mathbf{R}^B \rightarrow \mathbf{R}^B$.

The same simplifications described in Section 2.1 for scalar denoisers apply to our treatment of block denoisers as well. Namely, for $y \in \mathbf{R}^B$, we have $\eta_t(y) = \eta(y; \tau, \sigma_t)$ with σ_t the effective noise level at t -th iteration. The scaling relation $\eta(y; \tau, \sigma_t) = \sigma_t \eta(y/\sigma_t; \tau, 1)$ allow us to focus on the case $\sigma_t = 1$, denoted by $\eta(y; \tau) \equiv \eta(y; \tau, 1)$.

The AMP algorithm stays the same as defined by Eqs. (1.2) to (1.4) in this block-separable case except for calculation of the scalar Ω_t . This is given by

$$\Omega_t = \text{Ave}_m \text{div}(\eta(\text{block}_m(y^t), \tau^{t-1})).$$

Here Ave_m denotes uniform average over the index m , i.e. $\text{Ave}_m f(m) \equiv M^{-1} \sum_{i=1}^M f(m)$, and div is the divergence operator. Namely, if Jf is the Jacobian of the mapping $f : \mathbf{R}^B \mapsto \mathbf{R}^B$, $\text{div}(f) \equiv \text{Tr}(Jf)$. In the case $B = 1$ these formulas reduce to the ones for scalar AMP that we have already seen.

Firm Thresholding with Minimax Tuning							
ε	Pred	off.1000	off.2000	ci.1000.lo	ci.1000.hi	ci.2000.lo	ci.2000.hi
0.01	0.0550	0.0057	0.0040	0.0054	0.0060	0.0038	0.0043
0.02	0.0960	0.0065	0.0039	0.0062	0.0069	0.0036	0.0041
0.05	0.1920	0.0064	0.0047	0.0060	0.0068	0.0043	0.0050
0.10	0.3170	0.0067	0.0050	0.0063	0.0071	0.0047	0.0054
0.15	0.4190	0.0070	0.0049	0.0065	0.0074	0.0045	0.0052
0.20	0.5070	0.0068	0.0052	0.0064	0.0073	0.0048	0.0055
0.25	0.5840	0.0069	0.0055	0.0064	0.0073	0.0051	0.0058
AMP - Minimax Nonlinear Shrinker							
ε	Pred	off.1000	off.2000	ci.1000.lo	ci.1000.hi	ci.2000.lo	ci.2000.hi
0.01	0.0530	0.0075	0.0052	0.0072	0.0078	0.0050	0.0055
0.05	0.1840	0.0126	0.0099	0.0122	0.0130	0.0096	0.0102
0.10	0.3020	0.0183	0.0164	0.0178	0.0188	0.0160	0.0167
0.15	0.3980	0.0147	0.0127	0.0142	0.0151	0.0124	0.0131
0.20	0.4800	0.0148	0.0114	0.0143	0.0152	0.0110	0.0118
0.25	0.5510	0.0155	0.0124	0.0151	0.0160	0.0120	0.0127
Soft Thresholding with Minimax Tuning							
ε	Pred	off.1000	off.2000	ci.1000.lo	ci.1000.hi	ci.2000.lo	ci.2000.hi
0.01	0.0610	0.0061	0.0045	0.0058	0.0064	0.0043	0.0048
0.02	0.1040	0.0065	0.0052	0.0062	0.0069	0.0049	0.0055
0.05	0.2030	0.0089	0.0072	0.0085	0.0093	0.0068	0.0075
0.15	0.4270	0.0117	0.0100	0.0112	0.0122	0.0096	0.0104
0.20	0.5110	0.0118	0.0100	0.0113	0.0124	0.0096	0.0104
0.25	0.5830	0.0127	0.0106	0.0122	0.0132	0.0102	0.0110

Table 2: Empirical Phase Transition studies for AMP algorithms based on three shrinkers. 1000 Monte Carlo repetitions at $N = 1000, 2000$. Column *Pred* corresponds to general formula (1.6), and is thought to accurate for large N . The columns *off.1000* and *off.2000* report values of the offset \widehat{PT} estimated by logistic regression, using Eqs. (2.7) and (2.8). (Note that offsets are systematically smaller at $N = 2000$ than at $N = 1000$, consistent with hypothesis that they vanish as $N \rightarrow \infty$.) Columns *ci.1000.lo*, *ci.1000.hi*, *ci.2000.lo*, *ci.2000.hi* give lower and upper endpoints of formal 95% confidence intervals for the offset.

3.1 Block Soft Thresholding

We first revisit the denoising problem for a single block.

Let $Y \in \mathbf{R}^B$ be a random vector obeying $Y = X + Z$ where $Z \sim \mathcal{N}(0, \sigma^2 I_B)$, independent of X , the object of interest. The block-soft thresholding operator $\eta^{soft}(\cdot; \tau)$ is the nonlinear shrinker $\eta^{soft}(Y; \tau) = Y \cdot (1 - \frac{\tau}{\|Y\|_2})_+$, where $\tau \geq 0$ is the threshold parameter. The case $B = 1$ reduces to traditional soft thresholding [DJ94]. Block thresholding has previously been considered by Tony Cai [Cai99] and also Hall, Kerkycharian and Picard [HKP98], although in specific ‘wavelet’ applications; our point of view is somewhat different.

Shrinkage makes sense in the situation where X is often likely to be zero or at least small. Let ν denote a probability distribution over \mathbf{R}^B and consider the dimension-normalized mean-squared error

$$\text{MSE}_B(\eta^{soft}(\cdot; \tau), \nu) = \frac{1}{B} \mathbb{E}\{\|X - \eta(Y; \tau)\|_2^2\},$$

where $X \sim \nu$, $Y \sim X + Z$. We will show that when X is block-sparse in the sense of being different from zero with small probability, shrinkage works well. Formally, we let $\mathcal{F}_{B,\varepsilon}$ denote the collection of distributions ν obeying

$$\nu(\{X = 0\}) \geq 1 - \varepsilon,$$

and define the minimax block soft threshold risk in dimension B by:

$$M_B(\varepsilon | \text{BlockSoft}) = \inf_{\tau} \sup_{\mathcal{F}_{B,\varepsilon}} \text{MSE}_B(\eta^{soft}(\cdot; \tau), \nu).$$

In Figure 5 we present graphs of $M(\varepsilon) = M_B(\varepsilon | \text{BlockSoft})$ as a function of ε . It is possible to prove the following structural properties: (i) $0 \leq M(\varepsilon) \leq 1$ (the upper bound follows from taking $\tau = 0$); (ii) $M(\varepsilon)$ is monotone increasing and concave (monotonicity is a consequence of $\mathcal{F}_{B,\varepsilon} \subseteq \mathcal{F}_{B,\varepsilon'}$ for $\varepsilon \leq \varepsilon'$, and concavity since any measure in $\mathcal{F}_{q\varepsilon_1 + (1-q)\varepsilon_2, B}$ can be written as convex combination of measures in $\mathcal{F}_{\varepsilon_1, B}$ and in $\mathcal{F}_{\varepsilon_2, B}$); (iii) $M(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0$; $M(\varepsilon) \rightarrow 1$ as $\varepsilon \rightarrow 1$. Recall that we are considering the dimension-normalized MSE. In terms of total-MSE, the unnormalized mean-squared-error, the last relation would be $\text{total-MSE}_B(\varepsilon) \rightarrow B$ as $\varepsilon \rightarrow 1$.

Associated with the minimax problem is also an optimal threshold value $\tau^*(\varepsilon | B)$, that we plot in Figure 3.1.

As $B \rightarrow \infty$ the minimax per-coordinate MSE has a well defined, and particularly explicit limit.

Lemma 3.1. *As $B \rightarrow \infty$, we have*

$$M_B(\varepsilon | \text{BlockSoft}) \rightarrow M_\infty(\varepsilon | \text{BlockSoft}) \equiv 2\varepsilon - \varepsilon^2.$$

Further, the dimension-normalized minimax threshold converges with increasing blocksize: $\tau^(\varepsilon | B) / \sqrt{B} \rightarrow \tau^*(\varepsilon | \infty) \equiv (1 - \varepsilon)$.*

This lemma is proved in Appendix C.1.

3.2 Block James-Stein

Starting from the perspective of ℓ_1 -penalized estimation, and compressed sensing, block soft thresholding seems very natural. However, the limiting relationship in Lemma 3.1 shows that this approach leaves room for improvement at large blocksizes. The most ambitious goal we could set,

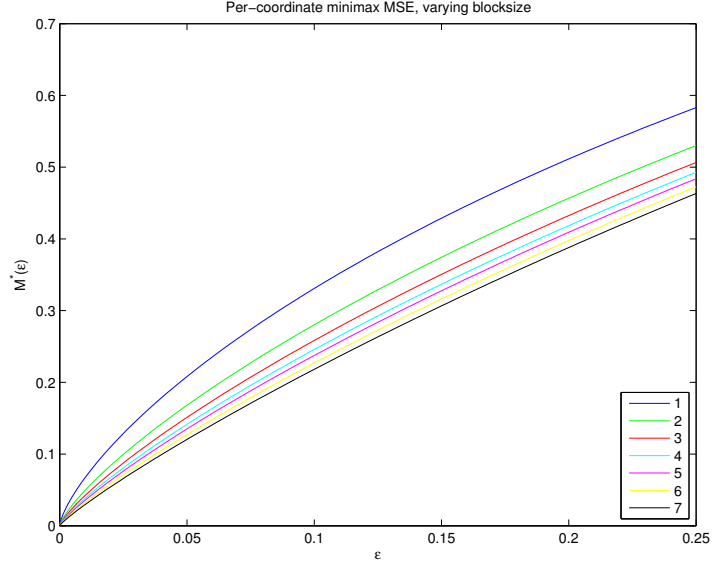


Figure 5: Minimax Soft-thresholding MSE as a function of ε , for $B = 1, 2, 3, 5, 7, 10$. MSE is normalized on a per-coordinate basis.

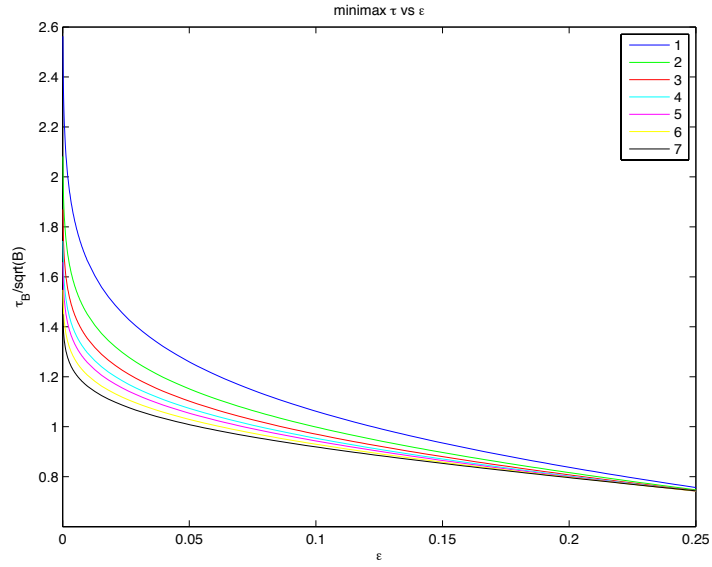


Figure 6: Minimax Threshold as a function of ε , for $B = 1, 2, 3, 5, 7, 10$. Threshold is normalized by \sqrt{B} .

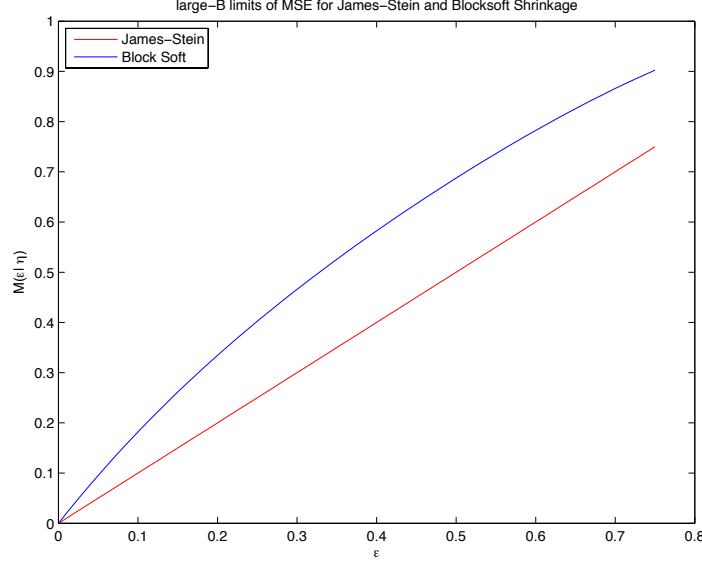


Figure 7: Limiting $B \rightarrow \infty$ per-coordinate MSE of BlockSoft and James-Stein shrinkers. The limiting James-Stein MSE coincides with the ideal MSE.

let's call it the *ideal MSE*, is the MSE achieved by a denoiser which utilizes a special oracle that tells us without error which blocks are zero and which are nonzero. The resulting ideal MSE is $M_B(\varepsilon|\text{Oracle}) = \varepsilon$. This is considerably smaller than $M_B(\varepsilon|\text{BlockSoft})$, even in the $B \rightarrow \infty$ limit characterized by Lemma 3.1. On the other hand the denoising/compressed sensing problems become easier as $B \rightarrow \infty$. Can we hope to achieve $M_B(\varepsilon|\text{Oracle})$?

To approach this ideal MSE in the large-blocksize limit, we propose to use the positive-part James-Stein shrinkage estimator [JS10]. On each block we propose the rule

$$\eta^{JS}(Y) = \left(1 - \frac{(B-2)}{\|Y\|_2^2}\right)_+ Y.$$

Analogously to block soft thresholding, this estimator shrinks to 0 blocks with small norms. However, the details are crucially different. The limiting $B \rightarrow \infty$ behavior is, in fact, ideal, and noticeably better than blocks soft shrinkage, as shown by Figure 3.2 and formally in the next lemma.

Lemma 3.2. *Let $M_B(\varepsilon|\text{JamesStein})$ denote the minimax MSE for η^{JS} over the class of ε -block sparse sequences. For $B > 2$:*

$$M_B(\varepsilon|\text{JamesStein}) \leq \varepsilon + \frac{2}{B}.$$

Proof. Consider temporarily the case where $Y = \mu + Z$ with $Z \sim \mathcal{N}(0, I_B)$, and $\mu \in \mathbf{R}^B$ nonrandom and known. A simple calculation shows that the optimal linear diagonal estimator η^{IDL} is given by

$$\eta^{IDL}(Y) = c(\mu) \cdot Y$$

where $c(\mu) = \|\mu\|_2^2/(\|\mu\|_2^2 + B)$. This estimator uses information about $\|\mu\|_2$ (which could only be supplied by an oracle) to choose the constant c as a function of $\|\mu\|_2$ which minimizes the mean-squared error among all rules of the estimators of form $c \cdot Y$. Note in particular that the per-coordinate MSE of this estimator is

$$\text{MSE}(\eta^{IDL}, \mu) = \frac{\|\mu\|_2^2}{\|\mu\|_2^2 + B},$$

and in particular

$$0 = \text{MSE}(\eta^{IDL}, \mu) \leq \sup_{\mu \in \mathbf{R}^B} \text{MSE}(\eta^{IDL}, \mu) = 1. \quad (3.2)$$

Define the ideal worst-case MSE by

$$M_B(\varepsilon|\eta^{IDL}) = \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} \mathbb{E}_\nu\{\text{MSE}(\eta^{IDL}, \mu = X)\}.$$

Applying (3.2), and keeping in mind that, under ν , $X = 0$ with probability at least $(1 - \varepsilon)$, and a bound 1 on the MSE of a nonzero block, we have

$$M_B(\varepsilon|\eta^{IDL}) = \varepsilon. \quad (3.3)$$

The oracle inequality [DJ95, Theorem 5] shows that for $B > 2$, and for every vector $\mu \in \mathbf{R}^B$, if $Y \sim \mathbf{N}(\mu, I_B)$, then

$$\text{MSE}(\eta^{JS}, \mu) \leq \text{MSE}(\eta^{IDL}, \mu) + \frac{2}{B}.$$

Combined with the previous display this proves the Lemma. \square

The argument in the proof leads in fact to a convenient expression for $M_B(\varepsilon|\text{JamesStein})$. Letting, with an abuse of notation, $\text{MSE}_B(\eta^{JS}, \mu) = \text{MSE}_B(\eta^{JS}, \delta_\mu)$ denote the per-coordinate MSE when the signal distribution is a Dirac delta $\nu = \delta_\mu$, we have

$$M_B(\varepsilon|\text{JamesStein}) = (1 - \varepsilon) \text{MSE}_B(\eta^{JS}, 0) + \varepsilon \max_{\mu} \text{MSE}_B(\eta^{JS}, \mu).$$

Now η^{JS} is known to be minimax for the unconstrained problem of estimating a non-sparse vector μ , i.e. $\max_{\mu} \text{MSE}_B(\eta^{JS}, \mu) = 1$ yielding

$$M_B(\varepsilon|\text{JamesStein}) = (1 - \varepsilon) \text{MSE}_B(\eta^{JS}, 0) + \varepsilon.$$

Therefore computing the minimax MSE for η^{JS} reduces to computing the single quantity $\text{MSE}_B(\eta^{JS}, 0)$, that can be estimated through numerical integration. A good approximation for large B is provided by the following formula $\text{MSE}_B(\eta^{JS}, 0) \approx B^{-1}(1 + 0.752/\sqrt{B})$. Hence we have

$$M_B(\varepsilon|\text{JamesStein}) \approx (1 - \varepsilon) \left\{ \frac{1}{B} + \frac{0.752}{B^{3/2}} \right\} + \varepsilon. \quad (3.4)$$

In the next section will use this formula in comparing the general prediction of Eq. (1.6) with the empirical results for the James-Stein AMP algorithm.

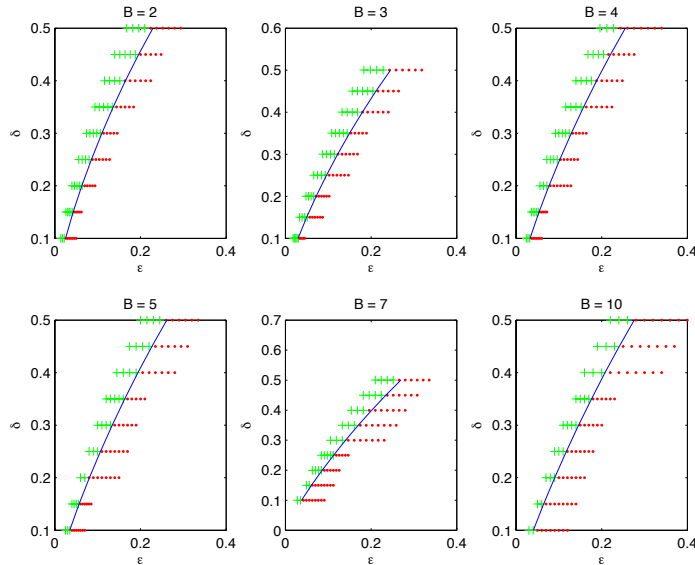


Figure 8: Results for block soft thresholding AMP with minimax threshold. Reconstruction is based on noiseless measurements $y = Ax_0$. Here $\delta = n/N$ is the undersampling fraction, and $\varepsilon = k/N$ is the sparsity measure. Red: less than 50% fraction of correct recovery. Green: greater than 50% fraction of successful recovery. Blue Curve: $\delta = M_B(\varepsilon|\text{BlockSoft})$.

3.3 Empirical Phase Transition Behavior

We now turn to the compressed sensing reconstruction problem whereby the block-sparse vector x_0 is reconstructed from observed data $y = Ax_0$ using the AMP algorithm. We want to test the hypothesis that Eq. (1.6) describes the phase transition of the two block shrinkage AMP algorithms, corresponding to the block soft thresholding, and block James-Stein.

We conducted a set of experiments similar to those described in Section 2.5 We constructed block-sparse signals at different undersampling and sparsity levels and ran tests of block thresholding AMP.

Our results show that the curve $\delta = M_B(\varepsilon|\text{BlockSoft})$ correctly separates two phases of performance: below this curve success in AMP recovery is atypical and above it is typical. Similarly, the curve $\delta = M_B(\varepsilon|\text{JamesStein})$ correctly describes the phase transition for block James-Stein shrinkage. The empirical results are presented in Figure 8 (for block soft thresholding) and Figure 9 (for block James-Stein).

4 Derivation of the Phase Transition by State Evolution

In this section we derive the basic relation (1.6) between minimax mean square error in denoising and the phase-transition boundary in the sparsity-undersampling plane. Our derivation is based on the state evolution formalism developed in [DMM09, DMM10, BM11a, DMM11].

One basic observation that is crucial for state evolution is that the mean squared error of the AMP reconstruction x^t at iteration t is practically non-random for large system sizes N , and has

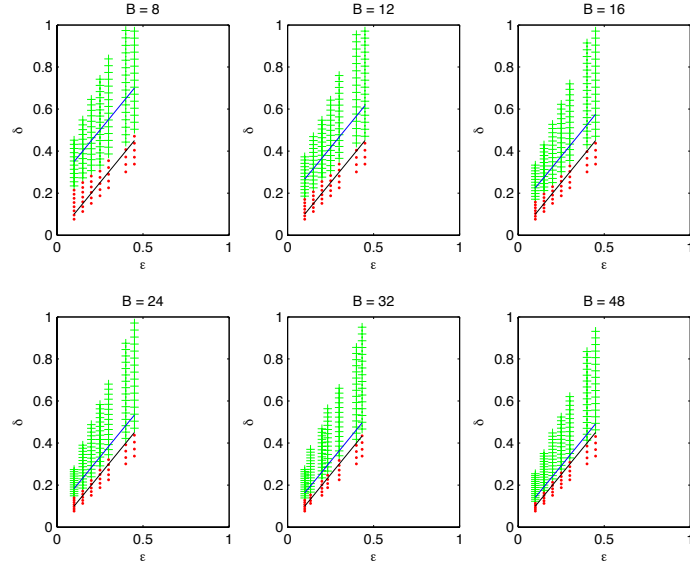


Figure 9: Results for block James-Stein AMP. Reconstruction is based on noiseless measurements $y = Ax_0$. Here $\delta = n/N$ is the undersampling fraction, and $\varepsilon = k/N$ is the sparsity measure. Red: less than 50% fraction of correct recovery. Green : greater than 50% fraction of successful recovery. Blue Curve: asymptotic (large- B) formula for $\delta = M_B(\varepsilon|\text{JamesStein})$ (3.4). Black Curve: $\delta = \varepsilon$ (lower bound on minimax risk of James-Stein Shrinkage).

a well-defined limit as $N \rightarrow \infty$. In particular, the limit

$$m_t = \lim_{N \rightarrow \infty} \frac{1}{N} \|x_0 - x^t\|_2^2,$$

exists almost surely (with $n \rightarrow \infty$ as well, while $n/N \rightarrow \delta$). Moreover, the evolution of m_t with increasing t is dictated by a formula $m_{t+1} = \Psi(m_t)$ which is explicitly computable, and defined below. We will use the term *state evolution* to refer both to the mapping $m \mapsto \Psi(m)$ and to the sequence $\{m_t\}_{t \geq 0}$ with appropriate initial condition. State evolution allows to determine whether AMP recovers the signal x_0 correctly, by simply checking whether $m_t \rightarrow 0$ as $t \rightarrow \infty$ (in which case the MSE vanishes asymptotically) or not. The latter problem does in turn reduce to a calculus problem.

The papers [DMM09, DMM10, DMM11], develop the state evolution framework for scalar shrinkers $\eta : \mathbf{R}^1 \mapsto \mathbf{R}^1$, and verified its predictions numerically for three specific examples (namely for the shrinkers Soft, SoftPos, and Cap). However, the heuristic argument presented in those papers was much more general. Indeed, [BM11a] proved that state evolution holds, in a precise asymptotic sense, for Gaussian measurement matrices A with iid entries and generic scalar shrinkers $\eta : \mathbf{R}^1 \mapsto \mathbf{R}^1$, under mild regularity assumptions.

Here we generalize this approach to vector shrinkers $\eta : \mathbf{R}^B \mapsto \mathbf{R}^B$. This framework covers all the shrinkers discussed in Sections 2 and 3, and yields a formal derivation of the main formula (1.6), under the assumption that indeed state evolution is correct in this broader context.

Before proceeding, let us remind the reader that we focused so far on denoisers applied to data $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$, whereby $\mathbf{Z} \sim \mathbf{N}(0, I_B)$ and hence the noise level is $\sigma = 1$. For the more general case of $\mathbf{Y} = \mathbf{X} + \sigma \mathbf{Z}$ where $\sigma \neq 1$, we recall the assumption of a scaling relation

$$\eta(\mathbf{Y}; \tau, \sigma) \equiv \sigma \cdot \eta(\sigma^{-1} \mathbf{Y}; \tau, 1). \quad (4.1)$$

In the next sections we will first introduce some basic notations and facts about state evolutions. Then we will derive the phase transition expression (1.6) by establishing first a lower bound and then a matching upper bound, both given by the minimax MSE.

4.1 State Evolution

The next definition provides the suitable generalization of the state evolution mapping to the present setting.

Definition 4.1. For given $\delta, \tau \geq 0$, $B \in \mathbf{N}$ and ν a probability distribution over \mathbf{R}^B , define the state evolution mapping $\Psi : \mathbf{R} \mapsto \mathbf{R}$ by

$$\Psi(m; \delta, \tau, \nu, B) \equiv \frac{1}{B} \mathbb{E}_\nu \left\{ \|X - \eta(X + \sqrt{m/\delta} Z; \tau, \sqrt{m/\delta})\|_2^2 \right\}.$$

Here, as before, X and Z are independent B -variate vectors, $X \sim \nu$, $Z \sim \mathbf{N}(0, I_B)$, and η is a given shrinker obeying (4.1). In other words, $\Psi(m; \delta, \tau, \nu, B)$ is the per-coordinate MSE at noise level $\sigma^2 \equiv m/\delta$.

Given implicit fixed parameters (δ, τ, ν, B) , state evolution is the one-dimensional dynamical system: $m_{t+1} = \Psi(m_t)$ starting from $m_0 = B^{-1} \mathbb{E}_\nu \|X\|_2^2$.

The fixed points of the mapping $m \mapsto \Psi(m)$ play of course a crucial role in the analysis of state evolution.

Definition 4.2. *The highest fixed point of the mapping $\Psi(\cdot) = \Psi(\cdot; \delta, \tau, \nu, B)$ is defined so: $\text{HFP}(\Psi) = \sup\{m : \Psi(m) \geq m\}$.*

The importance and applicability this notion is underscored by the next two observations. Here and below we say that a function $f : \mathbf{R}_+ \rightarrow \mathbf{R}$ is *starshaped* if $x \mapsto f(x)/x$ is decreasing.

Lemma 4.1. *Suppose any one of these three conditions holds:*

- $\Psi(m)$ is an increasing function of m with $m_0 > \text{HFP}(\Psi)$; or
- $\Psi(m)$ is an increasing starshaped function of m .
- $\text{HFP}(\Psi) = 0$.

Then state evolution converges to the highest fixed point:

$$\lim_{t \rightarrow \infty} m_t = \text{HFP}(\Psi).$$

The proof is a standard calculus exercise; we omit it.

Lemma 4.2. *The function $m \mapsto \Psi(m)$ is increasing starshaped for all of the following choices of the denoiser η , with the scaling convention (4.1):*

- *Soft Thresholding.*
- *Positive Soft Thresholding.*
- *Block Soft Thresholding.*
- *James-Stein Shrinkage.*

Proof. We give a proof only for Soft, Blocksoft, and James-Stein shrinkers; those denoisers are covariant under rotation. Explicitly, for any orthogonal matrix $Q \in \mathbf{R}^{B \times B}$, $\eta(Qy; \tau, \sigma) = Q\eta(y; \tau, \sigma)$. As a consequence, if $\tilde{\nu}$ is the probability measure obtained by averaging ν over rotations, we have $\Psi(m; \delta, \tau, \tilde{\nu}, B) = \Psi(m; \delta, \tau, \nu, B)$.

We can therefore, without loss of generality, consider measures ν that are themselves invariant under rotations. Any such measure can be written as convex combination of probability measures ω_μ that spread their mass uniformly on the sphere of radius $\mu > 0$. For a given fixed τ , let $R(\mu) = R(\mu; \tau)$ denote the risk function $R(\mu) = \text{MSE}(\omega_\mu; \tau, 1)$. From the scaling relation (4.1), we have the identity

$$\Psi(m) = \sigma^2 R(\mu/\sigma), \quad \sigma^2 = m/\delta.$$

We need to show that $\Psi(m)/m$ is nonincreasing in $m > 0$. However, this is just the same as asking if $R(\mu \cdot \sqrt{\delta/m})/\delta$ is nonincreasing in $m > 0$, and is equivalent to asking if $\mu \mapsto R(\mu)$ is monotone increasing in $\mu > 0$.

For each of the listed η , it is known that the corresponding $\mu \mapsto R(\mu)$ is strictly monotone increasing in $\mu > 0$. Details are provided in Appendix I. \square

4.2 State Evolution Phase Transition

Consider a collection $\mathcal{F}_{\varepsilon,B}$ of probability distributions over \mathbf{R}^B , indexed by $\varepsilon \in [0, 1]$. We will consider the following two properties:

1. $\mathcal{F}_{\varepsilon,B}$ is *nested*, i.e. $\mathcal{F}_{\varepsilon_1,B} \subseteq \mathcal{F}_{\varepsilon_2,B}$ whenever $\varepsilon_1 \leq \varepsilon_2$.
2. $\mathcal{F}_{\varepsilon,B}$ is *scale-invariant*, i.e. if $\nu \in \mathcal{F}_{\varepsilon,B}$ then any scaled version of ν is also in $\mathcal{F}_{\varepsilon,B}$.

In particular, the class $\mathcal{F}_{\varepsilon,B}$ of probability measures ν obeying the sparsity constraint $\nu(\{X = 0\}) \geq 1 - \varepsilon$ (introduced already above) is both nested and scale-invariant.

We temporarily use the notation $\text{HFP}(\delta, \tau, \nu) = \text{HFP}(\Psi(\cdot; \delta, \tau, \nu, B))$, and consider the minimax value:

$$\text{HFP}^*(\varepsilon, \delta) = \inf_{\tau \in \mathbf{R}_+} \sup_{\nu \in \mathcal{F}_{B,\varepsilon}} \text{HFP}(\delta, \tau, \nu).$$

Definition 4.3. For $\varepsilon \in [0, 1]$, define the state evolution phase transition as

$$\delta_{SE}(\varepsilon) \equiv \inf \{ \delta \geq 0 : \text{HFP}^*(\varepsilon, \delta) = 0 \}.$$

Note that $\text{HFP}^*(\varepsilon, \delta)$ is monotone decreasing as a function of δ , by definition of the state evolution mapping Ψ . It follows that $\text{HFP}^*(\varepsilon, \delta) = 0$ for $\delta > \delta_{SE}(\varepsilon)$. Further, by nestedness, it is monotone increasing as a function of ε , which implies that $\varepsilon \mapsto \delta_{SE}(\varepsilon)$ is monotone increasing. The rationale for this definition is that, for $\delta > \delta_{SE}(\varepsilon)$ and under any of the assumptions of Lemma 4.1, state evolution predicts that AMP will correctly recover the signal x_0 .

Theorem 4.1. Let $\mathcal{F}_{B,\varepsilon}$ be a nested, scale-invariant collection of probability distributions, and assume that the shrinker $\eta(\cdot; \tau, \sigma)$ obeys the scaling relation (4.1). Define the minimax MSE

$$M(\varepsilon) = \inf_{\tau} \sup_{\nu \in \mathcal{F}_{B,\varepsilon}} \frac{1}{B} \mathbb{E}_{\nu} \left\{ \|X - \eta(X + Z; \tau, 1)\|_2^2 \right\},$$

where $Z \sim \mathcal{N}(0, I_B)$ is independent of $X \sim \nu$. Then

$$\delta_{SE}(\varepsilon) = M(\varepsilon).$$

In order to prove this result, we shall first establish a more general fact. Given a distribution ν over \mathbf{R}^B , and for any $\tau \in \Theta$, we let

$$\delta_{SE}(\tau, \nu) \equiv \inf \{ \delta \geq 0 \text{ such that } \text{HFP}(\delta, \tau, \nu) = 0 \}. \quad (4.2)$$

The rationale for this definition is clear. Under the assumption that state evolution holds, for a signal x_0 sampled from block distribution ν , AMP (with tuning parameter τ) is guaranteed to reconstruct x_0 if and only if $\delta > \delta_{SE}(\tau, \nu)$.

Analogously, we let

$$M(\tau, \nu) = \sup_{\sigma > 0} \frac{1}{B\sigma^2} \mathbb{E}_{\nu} \left\{ \|X - \eta(X + \sigma Z; \tau, \sigma)\|_2^2 \right\}; \quad (4.3)$$

the normalized MSE for denoising with worst case signal. With these definitions we have the following.

Lemma 4.3. For any probability measure ν over \mathbf{R}^B , and any $\tau \in \Theta$, we have

$$\delta_{SE}(\tau, \nu) = M(\tau, \nu). \quad (4.4)$$

Proof. By definition

$$\begin{aligned} \delta_{SE}(\tau, \nu) &= \inf \left\{ \delta > 0 \text{ such that } \Psi(m; \delta, \tau, \nu, B) < m \text{ for all } m > 0 \right\} \\ &= \inf \left\{ \delta > 0 \text{ such that } \sup_{m>0} \frac{1}{m} \Psi(m; \delta, \tau, \nu, B) < 1 \right\} \\ &= \inf \left\{ \delta > 0 \text{ such that } \sup_{m>0} \frac{1}{mB} \mathbb{E} \{ \|X - \eta(X + \sqrt{m/\delta}Z; \tau, \sqrt{m/\delta})\|^2 \} < 1 \right\} \\ &= \inf \left\{ \delta > 0 \text{ such that } \sup_{\sigma>0} \frac{1}{\delta\sigma^2 B} \mathbb{E} \{ \|X - \eta(X + \sigma Z; \tau, \sigma)\|^2 \} < 1 \right\} \\ &= \inf \left\{ \delta > 0 \text{ such that } \frac{1}{\delta} M(\tau, \nu) < 1 \right\} = M(\tau, \nu). \end{aligned}$$

□

We are now in position to prove Theorem 4.1.

Proof of Theorem 4.1. Define $\delta_*(\varepsilon) \equiv \inf_{\tau \in \Theta} \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} \delta(\tau, \nu)$. We claim that $\delta_*(\varepsilon) = \delta(\varepsilon)$. Indeed choose $\delta \in [0, \delta(\varepsilon))$. Then by definition there exists $m > 0$ such that, for all $\tau \in \Theta$, there exists $\nu \in \mathcal{F}_{\varepsilon, B}$ with $\text{HFP}(\delta, \tau, \nu) > m$. Hence, for all $\tau \in \Theta$, there exists $\nu \in \mathcal{F}_{\varepsilon, B}$ with $\delta(\tau, \nu) \geq \delta$. This implies that, for all $\tau \in \Theta$, $\sup_{\nu \in \mathcal{F}_{\varepsilon, B}} \delta(\tau, \nu) > \delta$, i.e. $\delta_*(\varepsilon) \geq \delta$. Proceeding in the same way, it is immediate to prove that, for any $\delta \in [0, \delta_*(\varepsilon))$, we have $\delta(\varepsilon) > \delta$. Hence, we conclude that $\delta(\varepsilon) = \delta_*(\varepsilon)$ as claimed.

To conclude the proof, we note that, by Lemma 4.3, we have

$$\begin{aligned} \delta_*(\varepsilon) &= \inf_{\tau \in \Theta} \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} M(\tau, \nu) \\ &= \inf_{\tau \in \Theta} \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} \sup_{\sigma>0} \frac{1}{B\sigma^2} \mathbb{E}_\nu \left\{ \|X - \eta(X + \sigma Z; \tau, \sigma)\|_2^2 \right\} \\ &= \inf_{\tau \in \Theta} \sup_{\nu \in \mathcal{F}_{\varepsilon, B}} \frac{1}{B} \mathbb{E}_\nu \left\{ \|X - \eta(X + Z; \tau, 1)\|_2^2 \right\}, \end{aligned}$$

where the last equality follows from the scale invariant property of $\mathcal{F}_{B, \varepsilon}$. The last quantity is nothing but $M(\varepsilon)$. □

4.3 Non-Covergence of State Evolution

By Definition 4.3, for all $\delta < \delta_{SE}(\varepsilon)$, all probability distributions $\nu \in \mathcal{F}_{B, \varepsilon}$, and all initial conditions m_0 , state evolution converges to the zero error fixed point, namely $\lim_{t \rightarrow \infty} m_t = 0$. Viceversa, for $\delta > \delta_{SE}(\varepsilon)$, there exists $\nu \in \mathcal{F}_{B, \varepsilon}$, and an initial condition m_0 , such that $\lim_{t \rightarrow \infty} m_t = m_\infty > 0$. The reader might wonder whether this conclusion (nonconvergence to 0) also holds if we use the initial condition that is relevant for AMP, i.e. if, for any $\delta > \delta_{SE}(\varepsilon)$, there exists $\nu \in \mathcal{F}_{B, \varepsilon}$ such that, taking $m_0 = B^{-1} \mathbb{E}_\nu$, we have $\lim_{t \rightarrow \infty} m_t = m_\infty > 0$.

Lemmas 4.1 and 4.2 immediately imply that the answer is positive for soft thresholding, positive soft thresholding, block soft thresholding, and James-Stein shrinkage. It turns out that the answer is positive also for firm thresholding and the global minimax denoiser. In Appendix D we describe the argument for these cases.

5 Nonseparable Denoisers

To some readers, the application of formula (1.6) to a wide range of separable or block separable shrinkers may seem like a “minor” extension. We now show that the formula can be applied also to nonseparable shrinkers.

5.1 Monotone Regression

Let \mathcal{M} denote the cone of nondecreasing sequences:

$$\mathcal{M} = \{\mathbf{X} \in \mathbf{R}^N : x_{t+1} \geq x_t \text{ for all } t \in \{1, 2, \dots, N-1\}\}. \quad (5.1)$$

Let η^{mono} denote the shrinker of *monotone regression*, namely

$$\eta^{mono}(y) = \operatorname{argmin}_{c \in \mathcal{M}} \|y - x\|_2^2. \quad (5.2)$$

This shrinker is highly non-separable, as one can understand most clearly by studying the standard pool-adjacent-violators algorithm for solving it.

In the compressed sensing setting, one can implement exactly the AMP iteration (1.2)-(1.4) for this shrinker; let’s call the resulting procedure *monoreg AMP*.

To discuss the phase transition properties, we must generalize the notion of sparsity to *simplicity*: in this case, we say that a monotone sequence is k -simple if it has at most k points of increase. We can then consider a phase diagram with coordinates (ε, δ) , where $\varepsilon = k/N$ is the simplicity fraction and $\delta = n/N$ is the undersampling fraction. Once this convention is adopted, numerical experiments document a well-defined phase transition, however, it does not seem to follow any previously-documented theoretical curve. For example, it is definitely not the same as any of the transitions mentioned in the paper so far.

To apply formula (1.6), we need to calculate $M(\varepsilon|\text{MonoReg})$, which first of all requires clearly defining the notion! Our approach is explained in Appendix E.

5.2 Total Variation Minimization

Let η^{tv} denote the shrinker of *total variation penalized least-squares* [ROF92], also called fused LASSO [TSR⁺05]:

$$\eta^{tv}(\mathbf{Y}) = \operatorname{argmin}_{\mathbf{X} \in \mathbf{R}^N} \|\mathbf{Y} - \mathbf{X}\|_2^2 + \lambda \cdot TV(\mathbf{X}), \quad (5.3)$$

$$TV(\mathbf{X}) = \sum_{i=1}^{N-1} |x_{i+1} - x_i| \equiv \sum_{i=1}^{N-1} |\Delta x_i|. \quad (5.4)$$

Here we introduce the notation $\Delta x_i \equiv x_{i+1} - x_i$ for the discrete derivative. There is an extensive literature devoted to solving this denoising problem; see for example [VO96].

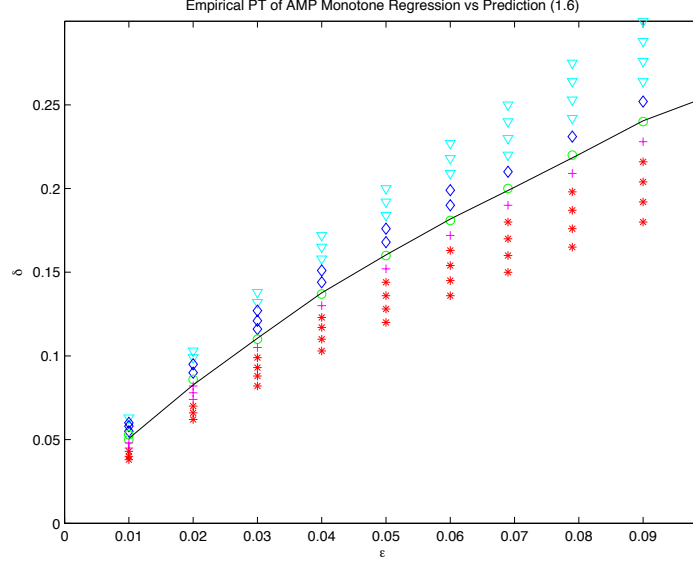


Figure 10: Results for monoreg AMP. Here $\delta = n/N$ undersampling fraction $\varepsilon = k/N$ generalized sparsity measure. Red – less than 50% fraction of correct recovery. Aqua – greater than 50% fraction of correct recovery. Curve – Minimax MSE curve $\delta = M(\varepsilon|\text{MonoReg})$.

In the compressed sensing setting, one can implement exactly the AMP iteration (1.2)-(1.4) for the TV denoiser. We will call the resulting procedure *TV AMP*.

To discuss the phase transition properties, we again generalize the notion of sparsity to *simplicity*: in this case, we say that a sequence is k -simple if it has at most k change points., and consider a phase diagram with coordinates (ε, δ) , where $\varepsilon = k/N$ is the simplicity fraction and $\delta = n/N$ is the undersampling fraction. To implement our formalism, we need to use TV-AMP with the tuning parameter λ that is optimal for the specified ε , which as in other cases requires a subsidiary computation.

Numerical experiments display a well-defined and seemingly novel phase transition.

To apply formula (1.6), we need to calculate $M(\varepsilon|TV)$; see Appendix F.

6 Discussion

6.1 Clarifications on AMP

The article [DMM09] defined the AMP algorithm for general scalar (separable) nonlinearities, and studied it for three specific cases: soft thresholding, positive soft thresholding and capping. The more general case was studied in [DMM10, BM11a],

However, the definition of the algorithm given in the introduction left out two details:

- To set the threshold level we need an estimate of the noise level. For sparse sequences, the

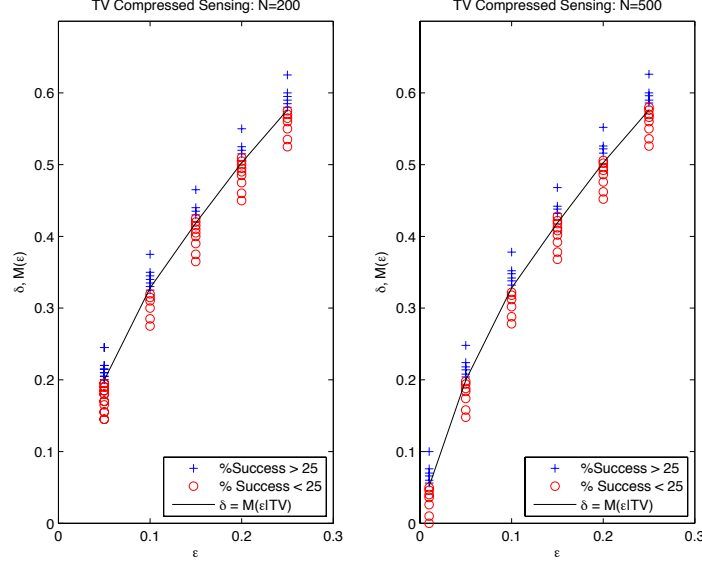


Figure 11: Results for TV AMP. $\delta = n/N$ undersampling fraction $\varepsilon = k/N$ generalized sparsity measure. Red – less than 25% correct recovery. Blue – greater than 25% fraction of correct recovery. Curve – Minimax MSE curve $\delta = M(\varepsilon|TV)$.

following estimate was proposed in [DMM09]:

$$\hat{\sigma}^t = \frac{1}{\Phi^{-1}(0.75)} \text{median}\{|z^t|\},$$

where $\Phi(z)$ is the normal distribution function. This is known as the 25% pseudo-variance in robust estimation; it is insensitive to a small fraction of large outliers.

- The formula for the working residual involves the scalar Ω^t ; in the scalar separable case $\eta : \mathbf{R} \mapsto \mathbf{R}$ it was defined in [DMM09] as

$$\Omega_t = \frac{1}{n} \cdot \sum_i \eta'(y^{t-1}; \tau^{t-1}).$$

Note the shift in indices: at each iteration we compute the next iteration's scalar. In the more general nonseparable case the formula is

$$\Omega_t = \frac{1}{n} \cdot \text{div} \eta(y; \tau^{t-1})|_{y=y^{t-1}}$$

where, for $f : \mathbf{R}^N \mapsto \mathbf{R}^N$, $\text{div} f$ is the divergence $\text{div} f = \sum_{i=1}^N \frac{\partial f_i}{\partial y_i}$. Note emphatically that although we are dealing with N -vectors, we divide by n .

6.2 Phase Transitions for other Algorithms

Formula (1.6) connects an algorithmic property – phase transitions of recovery algorithms – with a property from statistical decision theory – minimax mean squared error in denoising.

We have not explored here a further connection, joining the behavior of convex optimization with that of certain AMP algorithms. As proved in [BM11b], in the large system limit AMP with soft thresholding denoiser effectively computes the solution to

$$(P_{1,\lambda}) \quad \text{minimize} \quad \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1, \quad x \in \mathbf{R}^N.$$

for an appropriately calibrated $\lambda = \lambda(\tau)$. Similarly, AMP with positive soft thresholding denoiser is expected to solve

$$(P_{1,\lambda}^+) \quad \text{minimize} \quad \frac{1}{2}\|y - Ax\|_2^2 + \lambda\langle 1, x \rangle, \quad x \in \mathbf{R}_+^N,$$

where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product over \mathbf{R}^N , and 1 is the all-one vector. Finally, AMP with capping denoiser effectively solves

$$(P_\square) \quad \text{minimize} \quad \frac{1}{2}\|y - Ax\|_2^2, \quad x \in [0, 1]^N.$$

It follows that the AMP phase transitions are the same as the convex optimization phase transitions.
6

More generally, consider a generalized reconstruction method of the form

$$(P_J) \quad \text{minimize} \quad \frac{1}{2}\|y - Ax\|_2^2 + \lambda J(x), \quad x \in \mathbf{R}^N. \quad (6.1)$$

All of the above examples can be expressed in this form for a suitable convex regularizer $J : \mathbf{R}^N \rightarrow \mathbf{R}$. To this we can associate an AMP algorithm, by using the denoiser $\eta^J(\cdot; \tau)$ in Eq. (1.2), (1.3), (1.4), whereby

$$\eta^J(y; \tau) \equiv \arg \min_{x \in \mathbf{R}^N} \left\{ \frac{1}{2}\|y - x\|_2^2 + \tau J(x) \right\}. \quad (6.2)$$

(we also let $\eta(\cdot; \tau, \sigma) = \eta(\cdot; \tau \cdot \sigma)$). We will refer to this algorithm as to AMP- J .

We then have the following general correspondence, which follows immediately by writing the stationarity condition of problem $P_J = P_J(\lambda)$ and the fixed points of AMP- J .

Proposition 6.1. *Any fixed point x^∞ of AMP- J with fixed point parameters $\tau_\infty, \Omega_\infty, \sigma_\infty$ corresponds to a stationary point of $P_J(\lambda)$ with λ given by*

$$\lambda = \tau_\infty \sigma_\infty (1 - \Omega_\infty), \quad \Omega_\infty = \frac{1}{n} \operatorname{div} \eta(y; \tau^\infty)|_{y=y^\infty} \quad (6.3)$$

In particular, if the regularizer J is convex, then fixed points correspond to minimizers.

⁶ Previously, phase transitions for such convex programs were proven [Don06, DT09a, DT10b] and computed by techniques of combinatorial geometry; these were identified with phase transitions in random combinatorial structures, namely the neighborliness of face lattices of random polytopes. Thus, formula (1.6) connects fundamental problems of minimax decision theory to fundamental problems in high dimensional combinatorial geometry.

Notice that, for noiseless compressed sensing reconstruction, it is natural to consider the limit $\lambda \rightarrow 0$ of the above problem, corresponding to

$$\begin{aligned} & \text{minimize} && J(x) && x \in \mathbf{R}^N, \\ & \text{subject to} && y = Ax. \end{aligned}$$

We will next exemplify this correspondence for the examples examined in this paper. In the block-structured case of noiseless compressed sensing, we can compare blockSoft AMP to the following convex optimization problem:

$$\begin{aligned} (P_{2,1}) \quad & \text{minimize} && \sum_{m=1}^M \|block_m(x)\|_2, \\ & \text{subject to} && y = Ax. \end{aligned}$$

The norm $\|\mathbf{X}\|_{2,1} = \sum_{m=1}^M \|block_m(x)\|_2$ is called the mixed $\ell_{2,1}$ norm. In the block thresholding tests of Section 3, we also considered the empirical phase transition behavior of $(P_{2,1})$. Figure 12 verifies that the phase transition occurs around the location predicted by (1.6), just as with block soft thresholding AMP.

In the non-separable case of noiseless compressed sensing, we can compare monoReg AMP to the following convex optimization problem:

$$\begin{aligned} (P_{\text{mono}}) \quad & \text{minimize} && \langle 1, \Delta x \rangle \\ & \text{subject to} && y = Ax, \quad \Delta x \geq 0, \end{aligned}$$

where $\Delta x = (\Delta x_1, \dots, \Delta x_N)$, $\Delta x_i = (x_{i+1} - x_i)$. Figure 13 verifies that the phase transition occurs around the location predicted by (1.6), just as with monoReg AMP.

Finally, we can compare TV AMP to the following convex optimization problem:

$$\begin{aligned} (P_{tv}) \quad & \text{minimize} && TV(x), \\ & \text{subject to} && y = Ax, \end{aligned}$$

where again $TV(x) \equiv \sum_{i=1}^{N-1} |\Delta x_i|$. Figure 14 verifies that the phase transition occurs at the location predicted by (1.6), just as with TV AMP.

6.3 Interpretation as Penalized Optimization for Separable Denoiser

Scalar soft thresholding η^{soft} can be interpreted as the solution of a penalized optimization problem

$$(P_J) \quad \text{minimize} \quad \frac{1}{2}(y - x)^2 + J(x),$$

where $J(x) = \tau|x|$ (thus, we defined J to absorb the factor τ).

It turns out that a similar interpretation can be given for any scalar denoiser (and hence for any separable denoiser as well). In particular, the minimax and firm thresholding rules η^{all} and $\eta^{firm}(\cdot; \tau_1, \tau_2)$ are optimizers of *nonconvex* penalization schemes.

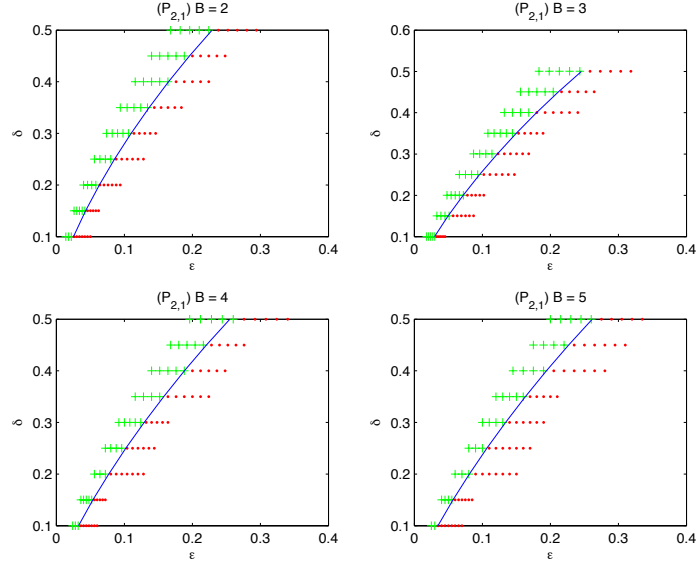


Figure 12: Results for $(P_{2,1})$. Here $\delta = n/N$ is the undersampling fraction and $\varepsilon = k/N$ is the sparsity measure. Red – less than 50% probability of correct recovery. Green – greater than 50% probability of correct recovery. Used Sedumi solver. Curve separating success from failure is the Minimax MSE curve $\delta = M(\varepsilon|\text{BlockSoft})$.

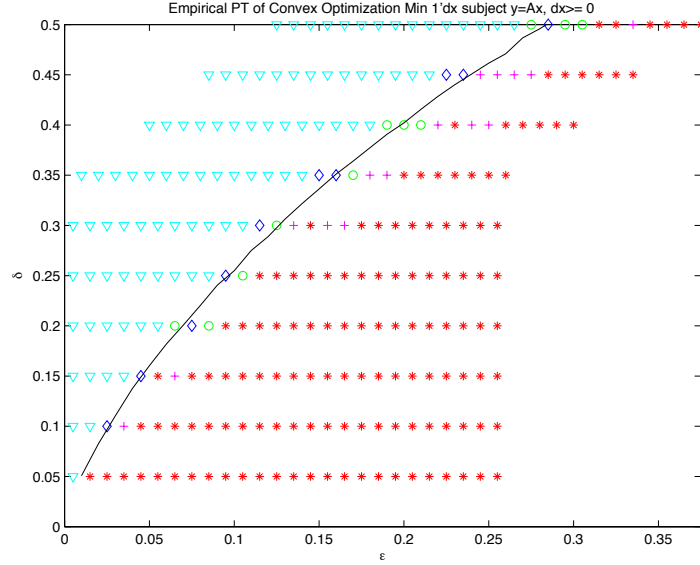


Figure 13: Results for (P_{mono}) . Here $\delta = n/N$ is the undersampling fraction and $\varepsilon = k/N$ is the generalized sparsity measure. Red – less than 50% fraction of correct recovery. Aqua – greater than 50% fraction of correct recovery. Curve – Minimax MSE curve $\delta = M(\varepsilon|\text{MonoReg})$. Compare Figure 10.

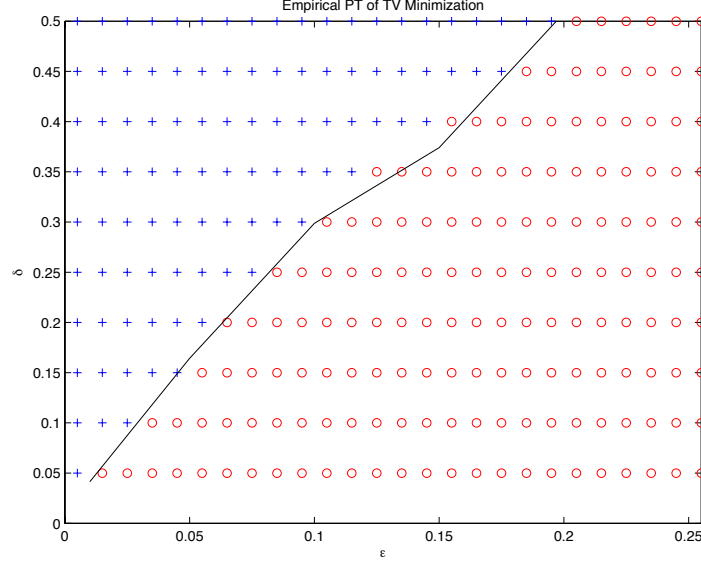


Figure 14: Results for (P_{TV}) . Here $\delta = n/N$ is the undersampling fraction and $\varepsilon = k/N$ is the generalized sparsity measure. Red – less than 25% fraction of correct recovery. Blue – greater than 25% fraction of correct recovery. Curve – Minimax MSE curve $\delta = M(\varepsilon|TV)$. Compare Figure 11.

We can derive the implied penalty $J(\cdot)$ from the nonlinearity $\eta(\cdot)$ by observing that $x + J'(x) = y$ at the solution $x = \eta(y)$. Defining the residual $\Delta(y) = y - \eta(y)$, and noting that $\eta(y) + \Delta(y) = y$, we observe that $\Delta(y)$ and $J'(x)$ are related through the change of variables

$$J'(\eta(y)) = \Delta(y).$$

In other words,

$$J(x) = \int_0^x \Delta(\eta^{-1}(u)) du.$$

Figure 15 shows the implied penalties for minimax firm and global minimax shrinkage, respectively. In both cases, the optimal penalizations are nonconvex, in accord with the commonly-held belief that nonconvex penalization is “superior” to ℓ_1 penalization [CY08, KXAH10]. However, the optimal penalization is seemingly very close to the ℓ_1 penalization, so that the prejudice in favor of nonconvex penalization must be re-examined.⁷

An implied penalty can also be derived for block James-Stein shrinkage η^{JS} . Due to the covariance under rotation, the corresponding penalty only depends on the modulus of $x \in \mathbf{R}^B$. Figure 16 shows the implied penalty $J(s|B)$ as a function of $s = \|x\|_2$. Namely, the penalty $J(\cdot|B) : \mathbf{R}_+ \rightarrow \mathbf{R}$ is such that, for $y \in \mathbf{R}^B$,

$$\eta^{JS}(y) = \operatorname{argmin}_{x \in \mathbf{R}^B} \min \left\{ \frac{1}{2} \|y - x\|_2^2 + J(\|x\|_2|B) \right\}$$

⁷The minimax nonlinearity is somewhat complicated to implement – see Appendix B. Figure 15 suggests to us that among piecewise linear rules, close to the best phase transition behavior is obtained by a rule which obeys $\eta(y) < y - \text{const}$ for large y , a possibility that is not explored in the firm family, or in any other conventional proposals.

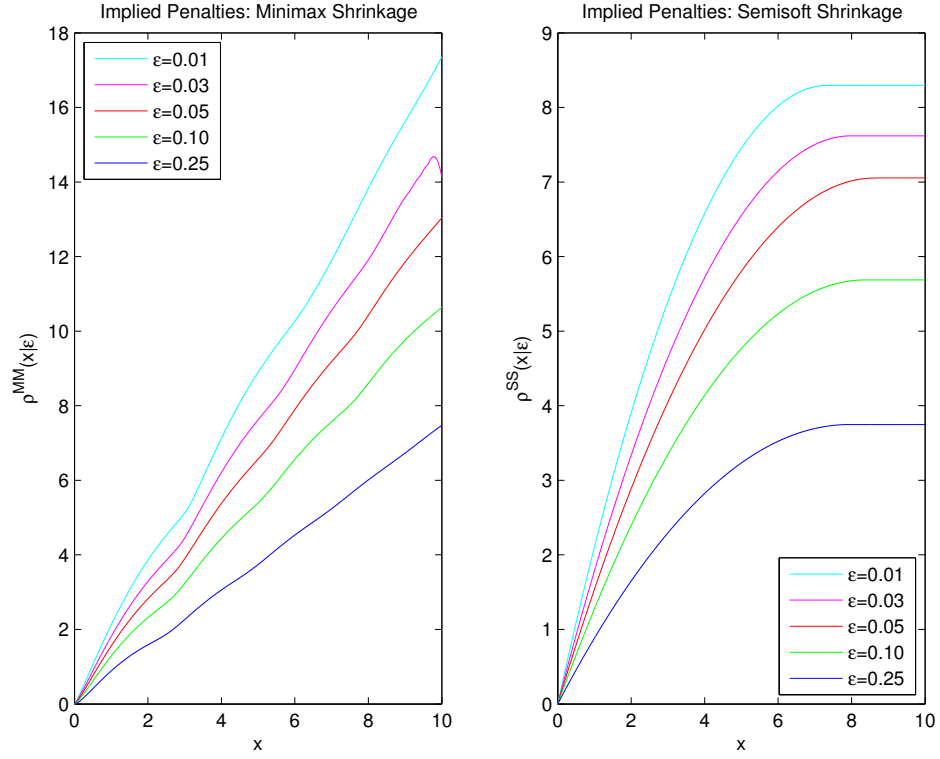


Figure 15: Implied Penalties $J(x)$ for two shrinkers: (a) Globally Minimax Shrinkage; (b) Minimax Firm Shrinkage; at $\epsilon \in \{0.01, 0.025, 0.05, 0.10, 0.25\}$ only results for positive arguments $x > 0$ are shown; results for negative arguments are obtained by symmetry $J(-x) = J(x)$.

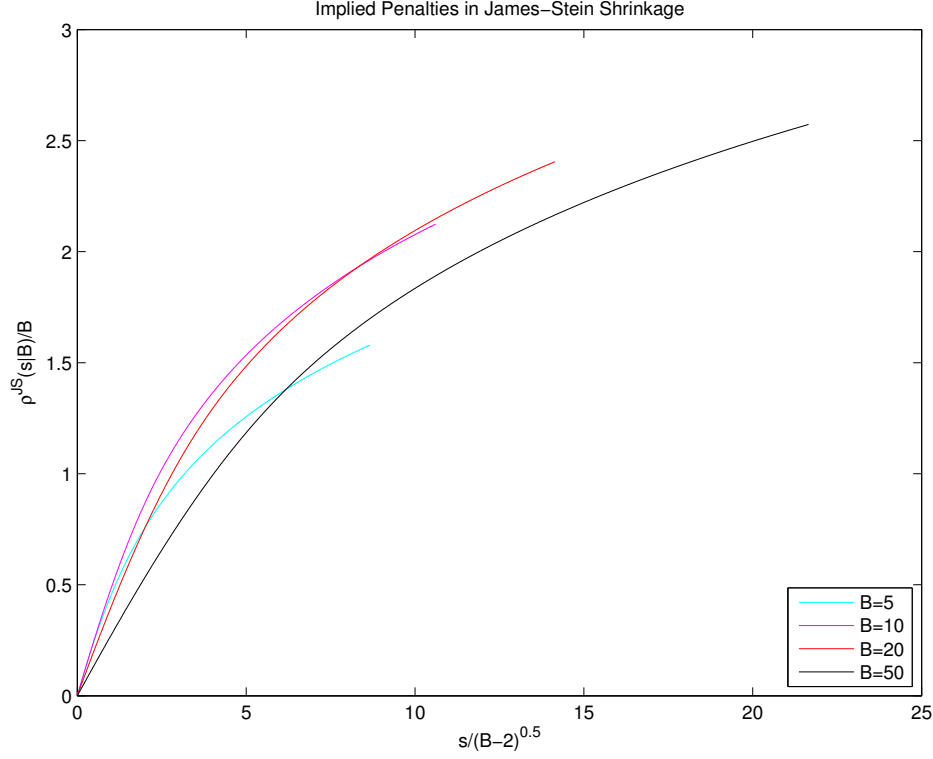


Figure 16: Implied penalties $J(s|B)$ for James–Stein positive-part estimator for block sizes $B \in \{5, 10, 20, 50\}$. Note that both axes are scaled with B dependent, so that all plots fit on a common scale. More precisely, the y -axis presents $J(s|B)/B$ and the x -axis presents $s/\sqrt{B-2}$.

coincides with the positive-part James–Stein estimator. The implicit penalization is again nonconvex.

6.4 Comparison to Traditional (δ, ρ) Phase Diagrams

In prior literature on phase transitions in compressed sensing, [DT10c, Don06, DT09a, BCT11, DMM09, DMM11], the authors considered a different phase diagram, based on variables δ and $\rho = \varepsilon/\delta$. The relation $\varepsilon = \rho \cdot \delta$ makes for a 1-1 relationship between the diagrams, so all information in the two diagrams can be presented in either format, however:

- The δ -axis is the y -axis here, whereas in the other cited work, it was the x -axis.
- The phase for successful recovery was *below* the $(\delta, \rho(\delta))$ curve in prior work, whereas here it is *above* the $(\varepsilon, \delta(\varepsilon))$ curve.

As a result, the ideal transition $\delta(\varepsilon|\text{Ideal}) = \varepsilon$ in the current framework corresponds to $\rho(\delta) = 1$ in the prior framework. In the prior framework $\rho = 1$ corresponds to *as many nonzeros as equations* in the the underlying system of equations, and so corresponds to a natural limit on any procedure.

Translating back and forth between these two frameworks can be instructive. Our result on James-Stein shrinkage achieving the ideal MSE $M(\varepsilon) = \varepsilon$ in the limit of large B can be restated as saying that block James-Stein can recover block sparse vectors with essentially as many nonzeros as measurements, in the limit of large blocksize.

6.5 Contributions of This Paper

We list eight contributions, beginning with the two most obvious:

- *Application of the AMP framework to a wider range of shrinkers $\eta(\cdot)$.* Here we have implemented and studied AMP algorithms that don't correspond to convex optimization (eg the Firm and Minimax scalar shrinkers) and also those that don't correspond to scalar separable nonlinearities: both the block separable case (block soft thresholding and block James-Stein shrinkage) and the general inseparable cases.
- *A formula for phase transitions of AMP algorithms.* As promised by the introduction, we see that formula (1.6) accurately describes the sparsity-indeterminacy combinations where AMP algorithms successfully recover a sparse signal from underdetermined measurements.
- *A formula predicting the phase transitions of many convex optimization problems.* Much work on compressed sensing establishes the possibility of recovery under sparsity by solving convex optimization problems. In most of the cases the trade-off between sparsity and undersampling is only known up to multiplicative constants. Quantitative guidance is provided by knowledge of the existence and location of phase transitions in the sparsity/indeterminacy plane. Unfortunately, considerable work was required to obtain those transitions for *one* convex optimization algorithm: ℓ_1 minimization [Don06, DT05, DT09a, DT10c]. The arguments needed to attack for example the block-sparsity case seemed to be quite different [Sto09]. In contrast, we have demonstrated here an approach which yields useful predictions in numerous cases, and possibly in many others.
- *Reconstruction approaches not based on convex optimization, with sharp guarantees.* We introduced three new AMP algorithms, based on Firm, Minimax, and James-Stein shrinkage, which are not derived from convex optimization. These methods have better phase transitions than the corresponding convex optimization problems, in their domains (e.g. Firm and Minimax outperform ℓ_1 minimization, while James-Stein outperforms block soft shrinkage for large B). To the extent that these methods can be viewed as solutions of optimization problems we show that the implied penalties are nonconvex. Nevertheless, these methods converge to the correct solution as long as we stay on the good side of the phase transition boundary; and in the interior of the success phase, these methods typically converge exponentially fast.
- *Limited benefit of nonconvex optimization for ordinary sparsity.* Within the class of scalar separable AMP algorithms, the best achievable phase transition is obtained by the minimax shrinker. Unfortunately the improvement in the transition is relatively small. This contradicts the hopes of many who have worked in this area, but not the actual results they have presented.
- *Existence of algorithms for the block sparse case approaching "Ideal" behavior.* While most attention in the group sparse case concerns block soft thresholding (and the corresponding

$\ell_2 - \ell_1$ penalized regression, a.k.a. group LASSO [YL10]), we show here that the phase transitions for block thresholding do not tend as $B \rightarrow \infty$ to the ideal transition, corresponding to as many nonzeros as equations. We also show that positive-part James-Stein shrinkage *does* tend to such an ideal limit.

- *Identification of Combinatorial Geometry Phase Transitions with Minimax Mean-Squared Errors.* The phase transitions for ℓ_1 optimization, and for positivity-constrained ℓ_1 optimization are determined by combinatorial geometry; see [DT09a]. By our general formula (1.6), these transitions are numerically the same as the minimax MSE in problems of scalar shrinkage, the empirical evidence given in the supplement of [DMM09], reinterpreted in the light of (1.6) shows this. In the appendix, we also prove analytically that a similar relation holds for another phase transition of combinatorial geometry: the phase transition for uniqueness of nonnegative sparse solutions to underdetermined linear equations happens at the same place as the minimax MSE for the positive-part estimator.
- *Calculation of the minimax MSE of monotone regression and total variation denoising.* We are not aware of any previous work deriving the minimax MSE of these denoising procedures under the condition of ε -sparse first differences. We give here a derivation and show that it agrees with the phase transition of both AMP and convex optimization algorithms.

Two conjectures flow naturally from this work:

- *State Evolution accurately describes the behavior of a wide range of AMP algorithms, for large system sizes N .* As we have seen, State Evolution allows one to derive (1.6). In the case of ℓ_1 penalization the correctness of state evolution as a description of AMP is proved by [BM11b]. Since formula (1.6) is apparently successful in some generality, this suggests the conjecture that state evolution applies much more generally than to the cases proven so far.
- *The AMP framework applies to a much wider range of shrinkers $\eta(\cdot)$ than proposed here.* If true, this would provide a general tool in compressed sensing. If one knew that a certain shrinker would be appropriate for the given type of content in the presence of direct noisy observations, then for the compressed sensing setting, one would simply AMP-ize it.

A Classical Cases

The general formula 1.6 was already derived in the Online Supplement to [DMM09], at in the cases of soft thresholding and positive soft thresholding. We review the arguments here since they provide a useful stepping stone for understanding more complicate cases (cf., e.g., Section 4).

Denoting by $\eta(\cdot; \tau)$ either soft of thresholding or positive soft thresholding with threshold τ , and taking the worst case over (respectively) $\nu \in \mathcal{F}_{1,\varepsilon}$ or $\nu \in \mathcal{F}_{1,\varepsilon,+}$ we have

$$\begin{aligned}
M(\varepsilon|\eta) &= \inf_{\tau \in \mathbf{R}_+} \sup_{\nu \in \mathcal{F}_{1,\varepsilon} / \nu \in \mathcal{F}_{1,\varepsilon,+}} \mathbb{E}\{(\eta(X + Z; \tau) - X)^2\} \\
&= \inf_{\tau \in \mathbf{R}_+} \sup_{\nu = (1-\varepsilon)\delta_0 + \varepsilon\delta_\mu} \mathbb{E}\{(\eta(X + Z; \tau) - X)^2\} \\
&= \inf_{\tau \in \mathbf{R}_+} \sup_{\mu \in \mathbf{R}_+} \left\{ (1-\varepsilon) \mathbb{E}[(\eta(0 + Z; \tau) - 0)^2] + \varepsilon \mathbb{E}[(\eta(\mu + Z; \tau) - \mu)^2] \right\}.
\end{aligned}$$

Here $X \sim \nu$ and $Z \sim \mathbf{N}(0, 1)$ are independent random variables. The reduction to two-point distributions follows from the representation of $\mathcal{F}_{1,\varepsilon}$ as a mixture of such two point distributions, along with linearity of expectation. Finally, the supremum is only over $\mu \geq 0$ because the support of ν is in \mathbf{R}_+ for $\nu \in \mathcal{F}_{1,\varepsilon,+}$ or because the risk is even in μ for the case of soft thresholding. Define

$$M(\mu, \tau; \varepsilon) = (1 - \varepsilon) \mathbb{E}[(\eta(0 + Z; \tau) - 0)^2] + \varepsilon \mathbb{E}[(\eta(\mu + Z; \tau) - \mu)^2].$$

Both for $\eta = \eta^{soft}$ the soft thresholding denoiser, for $\eta = \eta^{softpos}$ the positive soft thresholding denoiser, we have that $\mu \mapsto M(\mu, \tau; \varepsilon)$ is monotone increasing in $\mu \geq 0$. Hence, in both cases,

$$\sup_{\mu} M(\mu, \tau; \varepsilon) = \lim_{\mu \rightarrow \infty} M(\mu, \tau; \varepsilon) \equiv M(\infty, \tau; \varepsilon).$$

It is easy to see that

$$M(\infty, \tau; 1) = 1 + \tau^2.$$

We also have

$$M(0, \tau; 0) = \mathbb{E}\{\eta(Z; \tau)^2\} = \begin{cases} \int_{[-\tau, \tau]^c} (|z| - \tau)^2 \phi(z) dz & \text{soft thresholding,} \\ \int_{[\tau, \infty]} (z - \tau)^2 \phi(z) dz & \text{positive soft thresholding.} \end{cases}$$

Here and below $\phi(z) \equiv e^{-z^2/2}/\sqrt{2\pi}$ is the standard normal density and $\Phi(x) \equiv \int_{-\infty}^x \phi(x) dx$ is the normal distribution function. Since we seek $M(\varepsilon|\eta) = \inf_{\tau \in \mathbf{R}} M(\infty, \tau; \varepsilon)$, we look for a global minimum of $\tau \mapsto M(\infty, \tau; \varepsilon)$. Now

$$\frac{\partial}{\partial \tau} M(\infty, \tau; \varepsilon) = (1 - \varepsilon) \cdot c \cdot [-2\phi(\tau) + 2\tau(1 - \Phi(\tau))] + \varepsilon \cdot 2\tau,$$

where $c = 1$ for positive soft thresholding and $c = 2$ for soft thresholding. We conclude that the derivative vanishes where

$$L(\tau) \equiv \frac{\phi(\tau)}{\tau} - \Phi(-\tau) = \frac{1}{c} \cdot \frac{\varepsilon}{1 - \varepsilon}.$$

The function L is monotone decreasing on $(0, \infty]$, and so there is a unique solution $\tau^*(\varepsilon)$ of this equation. The minimax risk is then given by

$$M(\varepsilon|\eta) = M(\infty, \tau^*(\varepsilon); \varepsilon).$$

As a last ‘classical’ example, we consider the capping nonlinearity $\eta^{cap}(x) = \min(1, \max(0, x))$. Unlike all the other shrinkers considered in this paper, η^{cap} is not scale-invariant. To correctly interpret this setting, we introduce the positive part nonlinearity $\eta^+(x) = x1_{\{x>0\}}$, and consider $\nu \in \mathcal{F}_{1,\varepsilon,+}$. Define

$$M(\varepsilon|\eta^+) = \sup_{\nu \in \mathcal{F}_{1,\varepsilon,+}} \mathbb{E}\{(\eta^+(X + Z) - X)^2\},$$

where $X \sim \nu$ and $Z \sim \mathbf{N}(0, 1)$ are independent random variables. Defining as above $M(\mu; \varepsilon) = (1 - \varepsilon)\mathbb{E}\{(\eta^+(0 + Z) - 0)^2\} + \varepsilon \cdot \mathbb{E}\{(\eta^+(\mu + Z) - \mu)^2\}$, we again have $\sup_{\mu} M(\mu; \varepsilon) = M(\infty; \varepsilon)$ (notice indeed that $\eta^+(y) = \eta^{softpos}(y; \tau = 0)$). We then have the explicit formula

$$M(\infty; \varepsilon) = (1 - \varepsilon) \int_0^\infty z^2 \phi(z) dz + \varepsilon \int_{-\infty}^\infty z^2 \phi(z) dz$$

whence

$$M(\varepsilon|\eta^+) = M(\infty; \varepsilon) = \frac{1}{2}(1 - \varepsilon) + \varepsilon,$$

and in particular,

$$\lim_{\varepsilon \rightarrow 0} M(\varepsilon|\eta^+) = \frac{1}{2}.$$

Notice that, for all the other shrinkers considered in this paper, we instead have $\lim_{\varepsilon \rightarrow 0} M(\varepsilon|\eta) = 0$. So sparsity is of limited value for this shrinker.

The paper [DT10b] considered the problem of recovering a vector $x \geq 0$ from measurements $y = Ax$, where A is an $n \times N$ matrix with iid $\mathcal{N}(0, 1)$ entries, by solving the simple feasibility problem. The authors derived a curve separating a success phase, in which the feasibility problem has, with high probability, a unique solution, from a failure phase, in which the problem admits multiple feasible solutions. Their curve, derived in (δ, ρ) coordinates (See Section 6.4) can be translated into (ε, δ) terms to yield

$$\delta(\varepsilon|\text{PosOrthant}) = \frac{1}{2} + \varepsilon/2, \quad 0 < \varepsilon < 1.$$

We have therefore the following rigorously-proven case in which the general formula (1.6) indeed extends to yield equality between phase transitions in convex geometry and minimax risk in statistical decision theory.

Corollary A.1. *Let $\delta(\varepsilon|\text{PosOrthant})$ denote the position of the N -large phase transition for uniqueness of solution to $y = Ax$ subject to $x \geq 0$. Let $M(\varepsilon|\eta^+)$ denote the minimax risk of the positive-part estimator $\eta^+(y) = (y)_+$ over the class $\mathcal{F}_{1,\varepsilon,+}$ of sparse positive means. Then*

$$\delta(\varepsilon|\text{PosOrthant}) = \frac{1}{2}(1 + \varepsilon) = M(\varepsilon|\eta^+), \quad 0 < \varepsilon < 1.$$

Further [DT10b], derived a formal equivalence between the phase transition for the positive orthant problem just described, and the phase transition for the hypercube problem where one observes $y = Ax$, $x \in [0, 1]^N$ and the fraction of coordinates x_i in x such that $x_i \in (0, 1)$ is at most ε . Hence

$$\delta(\varepsilon|\text{Hypercube}) = \frac{1}{2}(1 + \varepsilon), \quad 0 < \varepsilon < 1.$$

Now define $\mathcal{F}_{1,\varepsilon,\square} \equiv \{\nu : \text{supp}(\nu) \in [0, 1], \nu(\{0, 1\}) \geq 1 - \varepsilon\}$, and define the *scale-invariant* minimax risk

$$M(\varepsilon|\eta^{cap}) \equiv \sup_{\nu \in \mathcal{F}_{1,\varepsilon,\square}} \sup_{\sigma \in \mathbf{R}_+} \frac{1}{\sigma^2} \mathbb{E}\{(\eta^{cap}(X + \sigma Z) - X)^2\}.$$

Notice that the supremum over σ could have been introduced for the previous examples – soft thresholding and positive soft thresholding – as well. However, for scale invariant families of distributions, the two definitions give the same minimax risk.

We then have

$$M(\varepsilon|\eta^{cap}) = M(\varepsilon|\eta^+) = \frac{1}{2}(1 + \varepsilon), \quad 0 < \varepsilon < 1.$$

Then general formula (1.6) is again rigorously valid.

Corollary A.2. *Let $\delta(\varepsilon|\text{Hypercube})$ denote the position of the N -large phase transition for uniqueness of solution to $y = Ax$ subject to $0 \leq x \leq 1$. Let $M(\varepsilon|\eta^{\text{cap}})$ denote the scale-invariant minimax risk of the capping nonlinearity $\eta^{\text{cap}}(y) = \max(0, \min(1, x))$ over the class $\mathcal{F}_{1,\varepsilon,\square}$ of distributions supported on $[0, 1]$. Then*

$$\delta(\varepsilon|\text{Hypercube}) = \frac{1}{2}(1 + \varepsilon) = M(\varepsilon|\eta^{\text{cap}}), \quad 0 < \varepsilon < 1.$$

B Calculation of Minimax MSE

In this Appendix we describe the calculation of the global minimax risk over the class $\mathcal{F}_{1,\varepsilon}$.

Throughout this section we let $\text{MSE}(\nu, \eta) \equiv \mathbb{E}\{(\eta(X + Z) - X)^2\}$ where $X \sim \nu$ and $Z \sim \mathcal{N}(0, 1)$ are independent random variables. Using arguments of Bickel [Bic81], and Casella and Strawderman [CS81] we have quite generally that, if ν is allowed to range over a weakly compact and convex class \mathcal{F} of probability distributions, then

$$\inf_{\eta} \sup_{\nu \in \mathcal{F}} \text{MSE}(\nu, \eta) = 1 - \inf_{\gamma \in \mathcal{F}} I(\gamma \star \nu),$$

where γ is the Gaussian measure $\gamma(dx) = \phi(x)dx$, $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$, $\gamma \star \nu$ denotes the convolution of measures, and I denotes the Fisher information. For a probability measure $\nu_f(dx) = f(x)dx$, with density f with respect to the Lebesgue measure this is defined as

$$I(\nu_f) = \int \frac{(f')^2}{f}(x) dx.$$

The distribution ν^* minimizing the Fisher information (if it exists) is called the least-favorable distribution. Using arguments of Bickel and Collins [BC83], we know that this distribution has the form of a mixture of point masses

$$\nu^* = \sum_{i=-\infty}^{\infty} \alpha_i \delta_{\mu_i},$$

where $\sum_i \alpha_i = 1$, $\alpha_i > 0$ and the sequence $\{\mu_i\}_{i \in \mathbf{Z}}$ has no accumulation points. By [Bic81] the minimax nonlinearity is then of the form

$$\eta^*(y) = y - \psi^*(y),$$

where ψ^* is the so-called score function

$$\psi^*(y) = \frac{d}{dy} \log f^*(y),$$

and f^* is the density of $\nu^* \star \gamma$.

Focusing now specifically on the class $\mathcal{F} = \mathcal{F}_{1,\varepsilon}$, without loss of generality we can assume that the μ_i are monotone increasing and that $\mu_0 = 0$, with $\alpha_0 = (1 - \varepsilon)$. A conjecture of Mallows [Mal78] states that in fact we may take

$$\mu_i = c \cdot i \quad \alpha_i = (1 - \varepsilon) \cdot \varepsilon^i / 2, \quad i \in \mathbf{Z} \setminus \{0\}.$$

In other words, the conjectured least favorable distribution has the form of a two-sided geometric distribution on a scaled copy of \mathbf{Z} . While this conjecture is thought to be false [BC83], we proceed as though it were asymptotically correct, i.e. a correct model for the behavior of μ_i and α_i for large $|i|$.

For estimating the minimax risk numerically, we chose a parameter $K = 50$ and assumed a generalized “Mallows form” for $|i| > K$. More precisely, we assume an equispaced grid, and geometrically decaying weights. This is a little more general than what Mallows proposed, having 3 total degrees of freedom (spacing, total weight and rate of decay), rather than one. For $-K \leq i \leq K$, we allow the parameters α_i and μ_i to vary freely. In this way we obtained a parametric family ν_θ of probability distributions, with parameter $\theta = (\epsilon, (\alpha_i), (\mu_i), c_0, c_1)$, with

$$\nu_\theta = (1 - \epsilon)\delta_0 + \epsilon \cdot \sum_{k=1}^K \alpha_i \cdot (\delta_{\mu_i} + \delta_{-\mu_i})/2 + \frac{c_0}{2} \sum_{k>K} \epsilon^k (\delta_{c_1 k} + \delta_{-c_1 k}).$$

Define

$$I^+ \equiv \min\{I(\gamma \star \nu_\theta) : \theta \in R^{K+3}\}.$$

The quantity I^+ can be estimated numerically, and provides an upper bound on I^* . In order to get a lower bound, we use Huber’s minimax theorem [Hub64, HR09], which tells us that

$$\frac{1}{I(\gamma \star \nu^*)} = \inf_{\psi} \sup_g \frac{\int \psi^2 g}{(\int \psi' g)^2}, \quad (\text{B.1})$$

where in the supremum g ranges over all densities $\gamma \star \nu$ with $\nu \in \mathcal{F}_{1,\epsilon}$ and ψ ranges over all functions differentiable in measure. Therefore, for any specific ψ let g_ψ denote a least-favorable density. We then have

$$I(\gamma \star \nu^*) \geq J(\psi, g) \equiv \frac{(\int \psi' g_\psi)^2}{\int \psi^2 g_\psi}.$$

Let ν^+ denote the optimum of the parametric optimization and let ψ^+ denote the corresponding score function

$$\psi^+(y) = \frac{d}{dy} \log f^+(y),$$

where $f^+ = \nu^+ \star \gamma$. Let g^+ denote a maximizing density g for $J(\psi^+, g)$; by Huber’s theory, this can be chosen to be two-point mixture $(1 - \epsilon)\gamma(\cdot) + \epsilon\phi(\cdot - \mu^+)$ where μ^+ is chosen to achieve the worst case value on the right side of (B.1) and set $I_- = J(\psi^+, g^+)$.

We have the bound $I_- \leq I^* \leq I^+$. Such lower and upper bounds are, however, concerning exact integrals with exact optimization. Numerically, we compute integrals and extrema over fine grids with at least 100 samples per unit of range, getting not I^+ and I_- but instead numerical approximations \tilde{I}^+ and \tilde{I}_- .

Table 3 presents some information about numerical approximation results, which may help the reader assess its accuracy.

Some minimizing distributions obtained in this way are shown in Figure 17; the mass points (μ_i) are displayed in Figure 18.

Our numerical results, showing that $\tilde{I}_- \approx \tilde{I}_+$ allows us to infer that the “Mallows form” is approximately correct. The “minimax nonlinearity” we actually apply in our estimation and compressed sensing experiments is:

$$\eta^+(y) = y - \psi^+(y).$$

ε	K	$1 - I^+$	$\text{MSE}^*(\psi^+, \varepsilon)$
0.01	2	0.0533	0.0533
0.05	2	0.1841	0.1841
0.10	2	0.3026	0.3026
0.15	2	0.3983	0.3984
0.20	4	0.4802	0.4803
0.25	4	0.5516	0.5516

Table 3: *Numerical Values of Lower Bound on Minimax MSE ($1 - I^+(\varepsilon)$) and of Upper Bound on Minimax MSE.* The upper bound is the numerically computed worst-case MSE of the corresponding decision rule η^+ . In these cases the bounds agree except possibly in the fourth decimal place, where the upper bound might be larger.

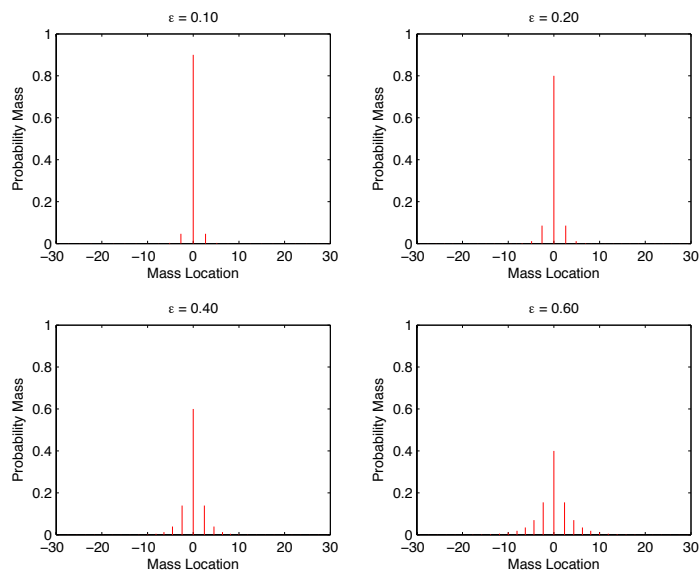


Figure 17: Least-Favorable distributions over the class $\mathcal{F}_{1,\varepsilon}$ obtained by numerically minimizing Fisher information.

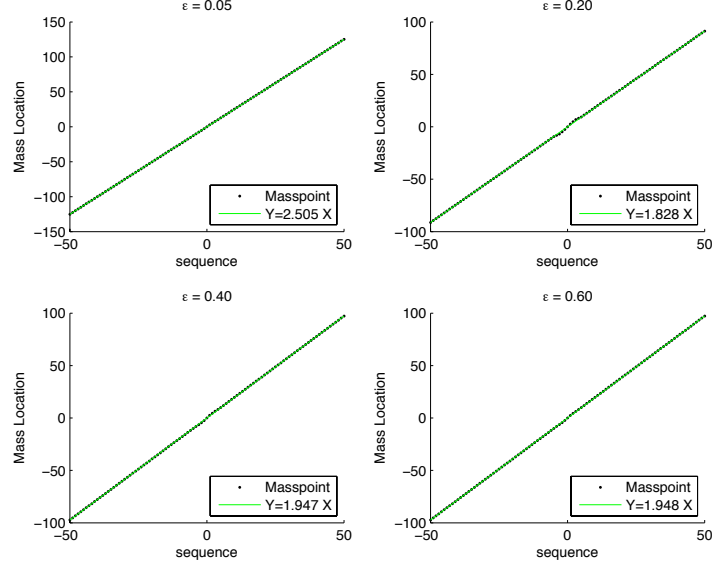


Figure 18: Locations of masspoints in approximate least-favorable distribution (μ_i) versus index i , at various ε , Note the approximate linearity; least-squares linear fit has slope parameter given in legend. The ‘wiggles’ away from the linear behavior occur near the origin.

C Calculation of Block Soft Thresholding Minimax MSE

The argument for the B -dimensional case is parallel to the scalar case (corresponding to $B = 1$) treated in Appendix A. Throughout this Appendix, we let $\eta(y; \tau) = \eta^{soft}(y; \tau)$ denote the block soft thresholding denoiser $\eta^{soft}(\cdot; \tau) : \mathbf{R}^B \rightarrow \mathbf{R}^B$. We then have

$$\begin{aligned}
M(\varepsilon|\eta) &= \inf_{\tau \in \mathbf{R}_+} \sup_{\nu \in \mathcal{F}_{B,\varepsilon}} \frac{1}{B} \mathbb{E}\{\|\eta(X + Z; \tau) - X\|^2\} \\
&= \inf_{\tau \in \mathbf{R}_+} \sup_{\nu=(1-\varepsilon)\delta_0+\varepsilon\delta_\mu; \mu \in \mathbf{R}^B} \frac{1}{B} \mathbb{E}\{\|\eta(X + Z; \tau) - X\|^2\} \\
&= \inf_{\tau \in \mathbf{R}_+} \sup_{\mu \in \mathbf{R}^B} (1 - \varepsilon) \cdot \frac{1}{B} \mathbb{E}\{\|\eta(0 + Z; \tau) - 0\|^2\} + \varepsilon \cdot \frac{1}{B} \mathbb{E}\{\|\eta(\mu + Z; \tau) - \mu\|^2\},
\end{aligned}$$

where $X \sim \nu$ and $Z \sim \mathbf{N}(0, 1)$ are independent, and the reduction to two-point distributions follows from the representation of $\mathcal{F}_{B,\varepsilon}$ as a mixture of such two point distributions, along with linearity of expectation. Defining

$$M_B(\mu, \tau; \varepsilon) \equiv (1 - \varepsilon) \cdot \frac{1}{B} \mathbb{E}\{\|\eta(0 + Z; \tau) - 0\|^2\} + \varepsilon \cdot \frac{1}{B} \mathbb{E}\{\|\eta(\mu + Z; \tau) - \mu\|^2\},$$

we note that $t \mapsto MSE_B(t \cdot \mu, \tau; \varepsilon)$ is monotone increasing in $t > 0$ and in fact depends only on $\|\mu\|_2$. Hence

$$\sup_{\mu \in \mathbf{R}^B} M_B(\mu, \tau; \varepsilon) = \lim_{\|\mu\| \rightarrow \infty} M_B(\mu, \tau; \varepsilon) \equiv M_B(\infty, \tau; \varepsilon).$$

It is easy to compute

$$M_B(\infty, \tau; 1) = 1 + \frac{\tau^2}{B}.$$

We also have

$$M_B(0, \tau; 0) = \frac{1}{B} \mathbb{E}\{\|\eta(Z; \tau)\|^2\} = \frac{1}{B} \mathbb{E}\{(\sqrt{X_{0,B}} - \tau)_+^2\} = \frac{1}{B} \int_{\tau^2}^{\infty} (x^{1/2} - \tau)^2 f_{0,B}(x) dx,$$

where X_B denotes a chi-squared variable on B degrees of freedom, and $f_{0,B}$ is the univariate density of a chi-squared variable on B degrees of freedom. The explicit form of the latter is $f_{0,B}(x) = x^{B/2-1} e^{-x/2} 2^{-B/2} \Gamma(B/2)^{-1}$. Define the (complementary) incomplete α -th moments of the chi-squared $I_{0,B}(\tau, \alpha) = \int_{\tau^2}^{\infty} x^\alpha f_{0,B}(x) dx$; these are expressible by incomplete moments of the Gamma function. Then $M_B(0, \tau; 0) = \{I_{0,B}(\tau, 1) - 2\tau I_{0,B}(\tau, 1/2) + \tau^2 I_{0,B}(\tau, 0)\}/B$.

We seek to evaluate

$$M_B(\varepsilon | \text{BlockSoft}) = \inf_{\tau \in \mathbf{R}_+} M_B(\infty, \tau; \varepsilon),$$

and note that

$$M_B(\infty, \tau; \varepsilon) = (1 - \varepsilon) M_B(0, \tau; 0) + \varepsilon M_B(\infty, \tau; 1).$$

In order to determine the minimum of $\tau \mapsto MSE_B(\infty, \tau; \varepsilon)$, we compute its first derivative and obtain

$$B \frac{\partial}{\partial \tau} M_B(\infty, \tau; \varepsilon) = (1 - \varepsilon) \cdot (-2I_{0,B}(\tau, 1/2) + 2\tau I_{0,B}(\tau, 0)) + \varepsilon \cdot 2\tau.$$

We conclude that the derivative vanishes where

$$L_B(\tau) \equiv \frac{I_{0,B}(\tau, 1/2)}{\tau} - I_{0,B}(\tau, 0) = \frac{\varepsilon}{1 - \varepsilon}.$$

The function L_B is monotone decreasing on $(0, \infty]$, and so there is a unique solution $\tau_B^*(\varepsilon)$. The minimax mean square error is given by

$$M_B(\varepsilon | \eta) = M_B(\infty, \tau_B^*(\varepsilon); \varepsilon).$$

C.1 Asymptotics for Large Blocksize

In this appendix we prove the limiting result in Lemma 3.1.

The scaling of the problem with B makes it natural to consider thresholds of the form $\tau_{c,B} \sim c \cdot \sqrt{B}$. We claim that the per-coordinate MSE with such a choice is

$$\begin{aligned} \lim_{B \rightarrow \infty} M_B(\infty, \tau_{c,B}; \varepsilon) &= (1 - \varepsilon) \cdot \lim_{B \rightarrow \infty} M_B(0, \tau_{c,B}) + \varepsilon \lim_{B \rightarrow \infty} M(\infty, \tau_{c,B}) \\ &= (1 - \varepsilon) \cdot (1 - c)_+^2 + \varepsilon \cdot (1 + c^2). \end{aligned} \tag{C.1}$$

Calculus shows that $(1 - \varepsilon)(c - 1)^2 + \varepsilon \cdot (1 + c^2)$ is minimized at $c^* = 1 - \varepsilon$. We conclude that

$$\lim_{B \rightarrow \infty} M_B(\varepsilon | \text{BlockSoft}) = (1 - \varepsilon)\varepsilon^2 + \varepsilon \cdot (1 + (1 - \varepsilon)^2),$$

which coincides with the statement of Lemma 3.1.

In order to complete the proof, we sketch the derivation of the claim C.1. Of the two limits, the easy one is $\lim_{B \rightarrow \infty} M_B(\infty, \tau_{c,B}) = 1 + c^2$. This is immediate because $M_B(\infty, \tau) = 1 + \tau^2/B$ in every dimension.

The limit $\lim_{B \rightarrow \infty} M_B(0, \tau_{c,B}) = (c-1)^2$ follows instead from the central limit theorem. Indeed, let $X_{0,B}$ denote a central chi-squared with B degrees of freedom. Its square root is the norm of a standard Gaussian random vector in dimension B , and is effectively about \sqrt{B} . Indeed by the central limit theorem $\sqrt{2}(\sqrt{X_{0,B}} - \sqrt{B}) \Rightarrow_D \mathcal{N}(0, 1)$ as $B \rightarrow \infty$. So as $B \rightarrow \infty$,

$$\begin{aligned} \lim_{B \rightarrow \infty} M_B(0, \tau_{c,B}) &= \lim_{B \rightarrow \infty} \frac{1}{B} \mathbb{E}\{(\sqrt{X_{0,B}} - \tau_{c,B})_+^2\} \\ &= \lim_{B \rightarrow \infty} \frac{1}{B} \mathbb{E}\{(\sqrt{B} + \frac{1}{\sqrt{2B}}Z - \tau_{c,B})_+^2\} \\ &= \lim_{B \rightarrow \infty} \mathbb{E}\{(1 - c + \frac{1}{\sqrt{2B}}Z)_+^2\} \end{aligned}$$

and the latter converges to $(1 - c)_+^2$ by dominated convergence. \square

D Non-Convergence of State Evolution

In this appendix we show non-convergence of state evolution for firm-thresholding AMP and globally minimax AMP below $\delta_{SE}(\varepsilon)$. More precisely, we show that, for $\delta < \delta_{SE}(\varepsilon)$ there exists a probability distribution ν such that the state evolution sequence $\{m_t\}_{t \geq 0}$, $m_0 = \mathbb{E}_\nu\{X^2\}$ does not converge to 0.

We begin by developing a lower bound that holds for all the shrinkers η studied in this paper. Let $\nu_{\varepsilon, \mu}$ denote the mixture $(1 - \varepsilon)\delta_0 + \varepsilon\omega_\mu$ where ω_μ is either (if $B = 1$), the Dirac mass at μ , or (if $B > 1$) the uniform measure on the unit ball of radius μ in \mathbf{R}^B . In both cases, let $\tau^*(\varepsilon)$ denote the minimax tuning for η . Here $\mu \in \mathbf{R}_+ \cup \{\infty\}$ can be either finite or infinite.

Definition D.1. Let $R(\mu) = R(\mu, \tau) = \text{MSE}_B(\eta(\cdot); \nu_{\varepsilon, \mu})$ denote the risk function of denoiser $\eta(\cdot; \tau)$ with tuning τ for the signal with distribution $\nu_{\varepsilon, \mu}$. We say that R is superquadratic on $[0, \mu_0)$ if, for any $\mu \in [0, \mu_0)$,

$$R(\mu) \geq \left(\frac{\mu}{\mu_0}\right)^2 \cdot R(\mu_0). \quad (\text{D.1})$$

The next result shows that superquadratic behavior of the risk function implies non convergence of state evolution for signal distribution $\nu_{\varepsilon, \mu}$.

Lemma D.1. (State Evolution non-Convergence) Fix any $\tau \in \Theta$, and assume there exists $\mu_0 > 0$ such that: (i) The risk function $R(\mu) = R(\mu; \tau)$ is superquadratic on $[0, \mu_0)$, (ii) $R(\mu_0) \geq \delta$, and (iii) $\delta \geq \varepsilon$.

Let $\Psi(m) = \Psi(m; \delta, \tau, \nu_{\varepsilon, \mu}, B)$. Then there is $m_{\text{fp}} > 0$ such that

$$m \geq m_{\text{fp}} \Rightarrow \Psi(m) \geq m_{\text{fp}}.$$

In particular, if $m_{\text{fp}} < m_0 = B^{-1}\mathbb{E}_{\nu_{\varepsilon, \mu}}\|X\|^2 = \varepsilon\mu^2/B$ then, for all $t \geq 0$,

$$m_t \geq m_{\text{fp}} > 0. \quad (\text{D.2})$$

Proof. By the scaling relation (4.1).

$$\Psi(m) \equiv \frac{1}{B} \mathbb{E}_{\nu_{\varepsilon, \mu}} \left\{ \|X - \eta(X + \sqrt{m/\delta}Z; \tau, \sqrt{m/\delta})\|_2^2 \right\} = \frac{m}{\delta} R\left(\frac{\mu}{\sqrt{m}}\right).$$

Define $m_{\text{fp}} = (\mu/\mu_0)^2$ and assume $m \geq m_{\text{fp}}$. This implies $\mu/\sqrt{m} \leq \mu_0$, and, since R is superquadratic by assumption,

$$\Psi(m) \geq \frac{m}{\delta} \left(\frac{\mu}{\sqrt{m}} \frac{1}{\mu_*} \right)^2 R(\mu_0) = \left(\frac{\mu}{\mu_0} \right)^2 = m_{\text{fp}},$$

which concludes the proof. \square

We also note that R need not itself be superquadratic, but merely dominate a superquadratic function, to get a similar conclusion.

Corollary D.1. *Suppose that, in the previous lemma, we merely assume that $R(\mu)$ is pointwise greater than a function $r(\mu)$ which is superquadratic on $[0, \mu_0]$ and where $r(\mu_0) = \delta$. The exact same conclusion follows.*

We resort to explicit verification of superquadraticity.

Numerical Observation. For $\eta \in \{\eta^F, \eta^{MM}\}$, let $\tau^*(\varepsilon|\eta)$ denote the minimax-tuning for sparsity ε and $\mu^*(\varepsilon|\eta)$ denote the least-favorable μ . The risk function $R(\mu) = R(\mu, \tau^*(\varepsilon|\eta)|\eta)$ is superquadratic on $(0, \mu^*(\varepsilon|\eta))$.

The underlying numerical study was as follows. For each shrinker, we took a fine grid of ε and at each fixed ε evaluated $R(\mu)$ on a fine grid of μ , checking the inequality (D.1). See Figures 19 and 20.

Finding. The State Evolution Phase Transition obeys $\delta_{SE}(\varepsilon|\eta) = M(\varepsilon|\eta)$ for $\eta \in \{\eta^F, \eta^{MM}\}$.

E Minimax MSE for Monotone Regression

The continuous curve in Figure 10 was generated by the straightforward device of generating random monotone objects with $\varepsilon = k/N$ fraction of change points. The “change heights” were chosen to be very large, and the change points were distributed uniformly at random. To understand the rationale for this choice, we describe an underlying theory. To the best of our knowledge the following evaluation of minimax MSE over ε -simple sequences is novel.

Throughout this section $\eta = \eta^{\text{mono}}$ denotes the monotone regression denoiser (5.2). For $\mu \in \mathcal{M} \subseteq \mathbf{R}^N$ a monotone vector, $\text{MSE}(\eta, \mu, \sigma)$ denotes the corresponding mean square error in noise with variance σ^2 , i.e.

$$\text{MSE}(\eta, \mu, \sigma) = \mathbb{E}\{\|\eta(\mu + \sigma Z) - \mu\|^2\}, \quad (\text{E.1})$$

where $Z = (Z_1, \dots, Z_n) \sim \mathbf{N}(0, I_{N \times N})$. We also use the shorthand $R(\mu) = \text{MSE}(\eta, \mu, 1)$ for the risk at μ . We use two principles:

- *Monotonicity of the risk function.* The function $t \mapsto R(t \cdot \mu)$ is monotone increasing in $t \in \mathbf{R}_+$.
- *Fragmentation.* The limiting risk $\lim_{t \rightarrow \infty} R(t \cdot \mu)$ decouples into a sum of risks of smaller problems for reconstructing signals with no internal change points.

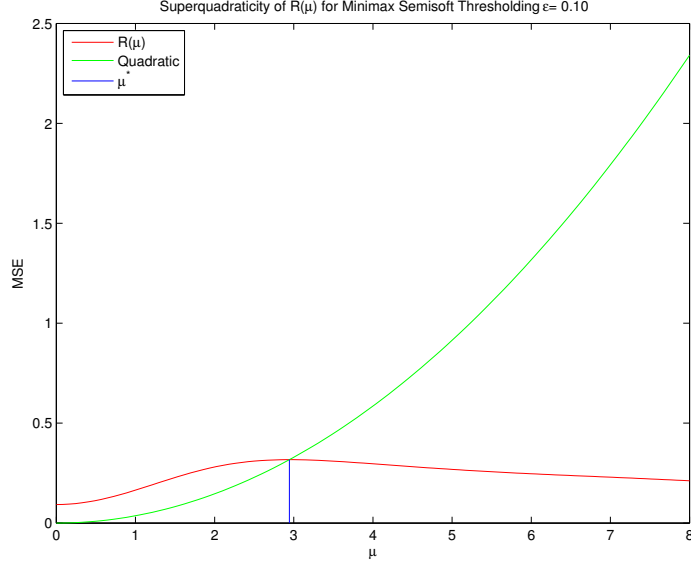


Figure 19: Numerical Verification of Superquadraticity for Firm Shrinkage. The green curve depicts the quadratic function $R(\mu^*) \cdot (\mu/\mu^*)^2$. The vertical line depicts the position of μ^* . The red curve depicts $R(\mu)$. The fact that the red curve is above the green curve throughout the interval $[0, \mu^*)$ verifies the superquadratic condition (D.1).

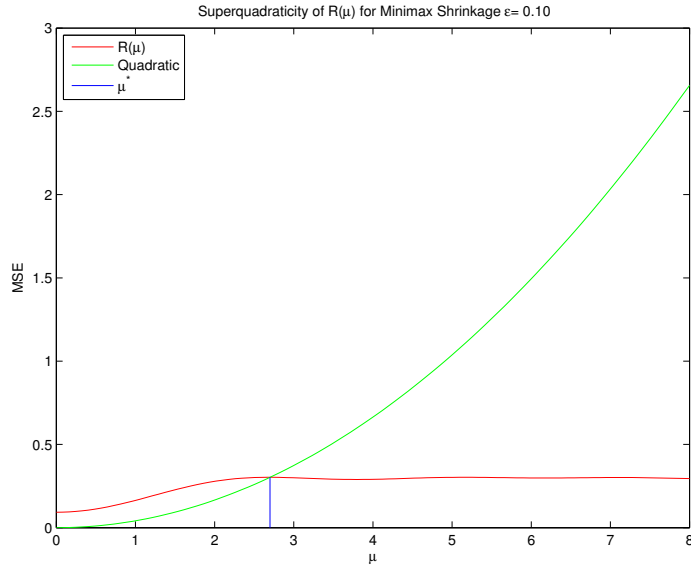


Figure 20: Numerical Verification of Superquadraticity for Minimax Shrinkage. The green curve depicts the quadratic function $R(\mu^*) \cdot (\mu/\mu^*)^2$. The vertical line depicts the position of μ^* . The red curve depicts $R(\mu)$. The fact that the red curve is above the green curve throughout the interval $[0, \mu^*)$ verifies the superquadratic condition (D.1).

The first principle is analogous to similar statements for positive soft thresholding, and the underlying reason is the same.

The second principle holds by the following argument. For a given monotone sequence $\mu\mathcal{M}$,

$$\text{MSE}(\eta^{mono}, t\mu, 1) = \text{MSE}(\eta^{mono}, \mu, 1/t) \cdot t^2.$$

This follows from the fact that the cone \mathcal{M} of monotone sequences is positive homogeneous, so we can equivalently consider scaling up the object μ or scaling down the noise σ and normalizing. Take the second viewpoint (i.e. the right hand side in the above equation). Introduce the optimization variable $v \equiv (x - \mu)/\sigma$. Since the noise vector is $z = (y - \mu)/\sigma$, the monotone regression problem is equivalent to the following

$$\begin{aligned} & \text{minimize} && \|v - z\|_2^2, \\ & \text{subject to} && \Delta v_i \geq -\frac{1}{\sigma} \Delta \mu_i, \text{ for all } i \in \{1, \dots, N-1\}, \end{aligned}$$

where we recall that $\Delta v_i = v_{i+1} - v_i$ is the discrete derivative. (Of course this problem does not provide an algorithm since it requires to know $\Delta \mu_i$, but here we are interested in it only for analysis purposes.)

As $\sigma \rightarrow 0$, all the constraints $\Delta v_i \geq -\frac{1}{\sigma} \Delta \mu_i$ for which $\Delta \mu_i > 0$ become irrelevant. Let I_0 denote the set of sites $i \in \{1, \dots, N-1\}$ where $\mu_{i+1} = m\mu_i$. We are naturally led to defining the following *localized* monotone regression problem

$$\begin{aligned} (Q_{lmono}) \quad & \text{minimize} && \|v - z\|_2^2, \\ & \text{subject to} && \Delta v_i \geq 0, \text{ for all } i \in I_0. \end{aligned}$$

Let $\eta^{lmono}(z; I_0)$ denote the solution of (Q_{lmono}) with data z, I_0 . When $z \sim \mathcal{N}(0, 1)$, we have the following limit in probability:

$$\lim_{\sigma \rightarrow 0} \frac{1}{\sigma} (\eta^{mono}(\mu + \sigma z) - \mu) = \eta^{lmono}(z; I_0).$$

As a consequence, defining $R^{loc}(I_0) = \mathbb{E}\{\|\eta^{lmono}(z; I_0)\|_2^2\}$, we have

$$\lim_{t \rightarrow \infty} \text{MSE}(\eta^{mono}, \mu, 1/t) \cdot t^2 = R^{loc}(I_0),$$

and therefore

$$\lim_{t \rightarrow \infty} R(t \cdot \mu) = R^{loc}(I_0).$$

In words, the worst-case risk of monotone regression is simply the local risk.

Now the set I_0 is a disjoint union $I_{0,1} \cup I_{0,2} \cup \dots \cup I_{0,K}$ of contiguous fragments, $I_{0,k} = [n_{k-1} + 1, \dots, n_k]$, where n_k is the last member of the k -th fragment. Mirroring this fragmentation, the problem (Q_{lmono}) separates into independent optimization problems. Define the *fragment optimization problem*

$$\begin{aligned} (Q_{lmono,I}) \quad & \text{minimize} && \sum_{i \in I} (z_i - v_i)^2, \\ & \text{subject to} && \Delta v_i \geq 0, \quad \text{if both } i, i+1 \in I, \end{aligned}$$

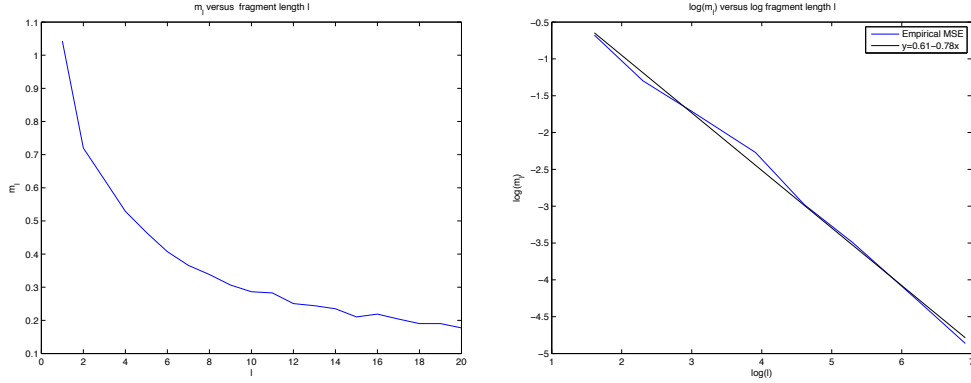


Figure 21: Fragment MSEs for Monotone Regression. Left panel: small ℓ ; Right panel: large ℓ . Left panel shows m_ℓ versus ℓ , $\ell = 1, \dots, 20$; Right panel shows $\log(m_\ell)$ versus $\log(\ell)$ for ℓ in 5, 10, 20, 50, 100, 200, 500, 1000. The m_ℓ were obtained by Monte-Carlo simulation.

where, if the fragment I is a singleton, the constraint disappears. Then the value of the overall optimization problem is the sum of the values of subproblems

$$\text{val}(Q_{lmono}) = \sum_{k=1}^K \text{val}(Q_{lmono, I_{0,k}}),$$

and the solution is the concatenation of subproblem solutions

$$\text{sol}(Q_{lmono}) = (\text{sol}(Q_{lmono, I_{0,1}}), \text{sol}(Q_{lmono, I_{0,1}}), \dots, \text{sol}(Q_{lmono, I_{0,K}})).$$

Clearly

$$R^{loc}(I_0) = \sum_{k=1}^K \mathbb{E}\{\|\text{sol}(Q_{lmono, I_{0,k}})\|_2^2\}.$$

On the other hand, $\mathbb{E}\|\text{sol}(Q_{lmono, I_{0,k}})\|_2^2$ clearly depends only on the cardinality of $I_{0,k}$. Hence, abuse notation by writing $Q_{lmono, \ell}$ for an optimization fragment of cardinality ℓ , and define

$$m_\ell = \ell^{-1} \cdot \mathbb{E}\{\|\text{sol}(Q_{lmono, \ell})\|_2^2\}.$$

Note in particular that $m_1 = 1$ and m_ℓ is monotone decreasing in ℓ . Figure 21, left panel, shows that for ℓ ranging from 1 to 20, the MSE is decaying rapidly. Figure 21, right panel, shows that for ℓ ranging from 5 to 1000, we have the approximate power law

$$\log(m_\ell) \approx 0.61 - 0.78 \log(\ell),$$

where the rational exponents $3/4$ and $4/5$ seem plausible for the underlying asymptotic relation.

Let π_ℓ denote the fraction of fragments having a given length, i.e. $\pi_\ell = K^{-1} \#\{k : |I_{0,k}| = \ell\}$. For the per-coordinate MSE we have

$$N^{-1} R^{loc}(I_0) = \sum_{\ell} \pi_\ell m_\ell \ell.$$

In short, the worst-case MSE depends only on the fragmentation characteristics and the behavior of the mean-squared solution of some standard problems.

In the simulations underlying Figure 10, the change points were placed uniformly at random, so the fragment lengths are approximately iid exponential random variables with mean $N/(k+1) \approx 1/\varepsilon$. Hence we used

$$M(\varepsilon|\text{MonoReg}) = \sum_{\ell} \pi_{\ell}^{\varepsilon} m_{\ell} \ell,$$

where the weights π_{ℓ}^{ε} follow the geometric distribution:

$$\pi_{\ell}^{\varepsilon} = (1 - \varepsilon)^{\ell-1} \varepsilon.$$

The preceding theoretical discussion makes clear why we chose to have very large change sizes at the change points (in practice change sizes larger than 3 are sufficient). It does not make clear why we considered change points distributed at random.

In fact we have only been considering $M(\varepsilon|\text{MonoReg})$, the minimax risk over objects with *randomly-distributed* change points. We could have considered instead $M^*(\varepsilon|\text{MonoReg})$, the minimax risk over objects with *least-favorable* change points.

Lemma E.1. *The least-favorable fragmentation is one which achieves the following supremum:*

$$M^*(\varepsilon|\text{MonoReg}) = \sup \left\{ \sum_{\ell} \pi_{\ell} m_{\ell} \ell : \sum_{\ell} \ell \pi_{\ell} = 1/\varepsilon \right\}.$$

Equivalently, the curve $(1/\varepsilon, M^(\varepsilon|\text{MonoReg}))$ is the least concave envelope of*

$$(\ell, m_{\ell} \ell), \quad \ell = 1, 2, \dots$$

Figure 22 shows that the concave envelope is apparently the interpolating linear spline. It follows that the least-favorable case has all fragments of equal length k/N . The geometric distribution is therefore (slightly) more favorable than the least-favorable change-point distribution.

F Minimax MSE for Total Variation Minimization

The curve $M(\varepsilon|TV)$ in Figure 11 was generated in a fashion paralleling the monotone regression case. We generated random piecewise constant objects with $\varepsilon = k/N$ fraction of change points. The “change heights” were chosen to be very large and randomly signed, and the change points were distributed uniformly at random.

The theoretical motivation is similar to the case of monotone regression, with additional features:

- the addition of a tuning parameter τ ,
- the addition of boundary conditions into the fragment optimization problems.

Skipping all the preliminaries, for $\ell > 1$ we define four fragment optimization problems, corresponding to different boundary conditions: no boundary conditions:

$$(Q_{tv,\ell,0,0}^{\lambda}) \quad \text{minimize} \quad \left\{ \frac{1}{2} \sum_{i=1}^{\ell} (z_i - v_i)^2 + \lambda \sum_{i=1}^{\ell-1} |v_{i+1} - v_i| \right\}.$$

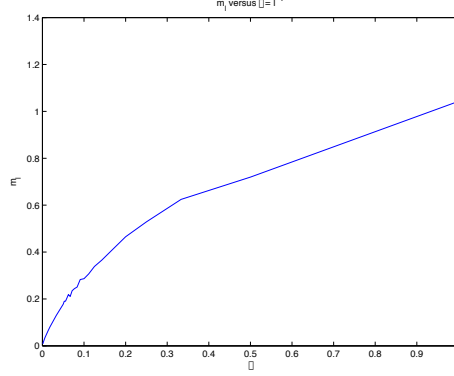


Figure 22: Fragment MSEs for Monotone Regression against reciprocal variable $\varepsilon = 1/\ell$. Plot combines the asymptotic law $m_\ell = \exp(0.61)\ell^{-0.78}$ for $\ell > 20$ with direct simulation for $\ell \leq 20$. Note the apparent concavity.

Even-parity boundary conditions:

$$(Q_{ltv,\ell,+,+}^\lambda) \quad \text{minimize} \quad \left\{ \frac{1}{2} \sum_{i=1}^{\ell} (z_i - v_i)^2 + \lambda \sum_{i=1}^{\ell-1} |v_{i+1} - v_i| + \lambda(v_1 + v_\ell) \right\}.$$

Odd parity boundary conditions:

$$(Q_{ltv,\ell,+,-}^\lambda) \quad \text{minimize} \quad \left\{ \frac{1}{2} \sum_{i=1}^{\ell} (z_i - v_i)^2 + \lambda \sum_{i=2}^{\ell} |v_i - v_{i-1}| + \lambda(v_1 - v_\ell) \right\}.$$

Boundary conditions at one end:

$$(Q_{ltv,\ell,0,+}^\lambda) \quad \text{minimize} \quad \left\{ \frac{1}{2} \sum_{i=1}^{\ell} (z_i - v_i)^2 + \lambda \sum_{i=2}^{\ell} |v_i - v_{i-1}| + \lambda v_\ell \right\}.$$

If $\ell = 1$, we have simply

$$\begin{aligned} (Q_{ltv,\ell,+,+}^\lambda) \quad & \text{minimize} \quad \left\{ \frac{1}{2} (z_1 - v_1)^2 + 2\lambda v_1 \right\}, \\ (Q_{ltv,\ell,0,+}^\lambda) \quad & \text{minimize} \quad \left\{ \frac{1}{2} (z_1 - v_1)^2 + \lambda v_1 \right\}. \end{aligned}$$

while

$$(Q_{ltv,\ell,+,-}^\lambda) \quad \text{minimize} \quad \frac{1}{2} (z_1 - v_1)^2,$$

and

$$(Q_{ltv,\ell,0,0}^\lambda) \quad \text{minimize} \quad \frac{1}{2} (z_1 - v_1)^2.$$

Define $m_{\ell,s}^\lambda = \ell^{-1} \cdot \mathbb{E}\{\|\text{sol}(Q_{\ell tv, \ell, s}^\lambda)\|_2^2\}$, for $s \in \{00, 0+, ++, +-\}$.

We let $F = (I, s)$ denote a fragment, where $I \subset \{1, \dots, N\}$ and $s \in \{00, 0+, ++, +-\}$.

For each sequence μ , the list of fragments $\{(I_k, s_k)\}$ is simply the list of pieces (subintervals of constancy) I_k , decorated as follows. First, suppose that piece I_k does not meet the endpoints 1 and N . If μ is a local extremum at this piece, we pick $s_k = ++$. If μ is not a local extremum, we pick $s_k = +-$. If I_k meets the endpoints 1, N , we have $s_k = 0+$ or 00 depending on whether I_k meets both endpoints or not. With these conventions, let $J = (F_1, \dots, F_K)$ denote a list of fragments; the local MSE conditional on the fragmentation is:

$$R^{loc}(J, \lambda) = \sum_{\ell, s} \pi_{\ell, s} m_{\ell, s}^\lambda \ell.$$

We then have the *random-changepoint minimax* MSE

$$M(\varepsilon|TV) = \min_{\lambda} \sum_{\ell, s} \pi_{\ell, s}^\varepsilon m_{\ell, s}^\lambda \ell$$

where $\pi_{\ell, s}^\varepsilon$ is the set of probabilities of fragment types induced by random changepoint assignment with a fraction of changepoints $\varepsilon = k/N$. We also have the *least-favorable changepoint* minimaxMSE

$$M^*(\varepsilon|TV) = \min_{\lambda} \sup \left\{ \sum_{\ell, s} \pi_{\ell, s} m_{\ell, s}^\lambda \ell : \varepsilon^{-1} = \sum_{\ell, s} \ell \cdot \pi_{\ell, s} \right\},$$

where the supremum is over all probabilities on the collection of decorated fragments, having mean length $1/\varepsilon$.

G Computational Details

Because this paper makes a heavy reliance on numerical evidence, we record here a number of details which may help the reader of the paper.

- *Environment.* All computations were done in Matlab running under Mac OS X. In the spirit of reproducible research, we intend eventually to make the code available for interested readers.
- *Numerical Computation of Minimax Risks.* For soft, hard, and firm thresholding, there are closed-form expressions for the MSE at a given mixture of point masses, in terms of $\int_a^b y^k \phi(y) dy$, where ϕ denotes the standard normal density. Compare Appendix A. For minimax thresholding, one must integrate on a fine grid; we used standard numerical integration routines in Matlab.

For block thresholding, there are closed-form expressions for the MSE at a given mixture of spherical masses in terms of incomplete beta functions. Compare Appendix C.

For the worst-case MSE over a class of probability measures, one must in principle maximize the MSE over all admissible measures. For the scalar- and block- separable smoothers considered this means maximization over two-point mixtures, $(1 - \varepsilon)\delta_0 + \delta_\mu$, which ultimately requires maximization of $\text{MSE}(\eta, \mu)$ across all μ . In some cases (Soft, BlockSoft, JamesStein),

the extremal μ is “at infinity” (a formula can be given in each case for the MSE “at infinity”); in other cases the least-favorable μ is finite and must be approximated by maximization over a fine grid. Generally speaking we performed grid searches by iterative subdivision: replace the interval in question into a finite grid of points (eg. 100), then evaluate by brute force on this grid and identify a subinterval with grid boundary points which contains the discrete optimum. Then replace the original search interval by the newly-found subinterval.

For the nonseparable cases, the process we followed for numerically approximating minimax MSE was explained in Appendices E and F.

- *Monte Carlo Simulations.* For scalar separable thresholding (soft, hard, and firm thresholding, and minimax shrinkage), we ran Monte Carlo simulations with the following parameters: 1000 Monte Carlo repetitions at each parameter combination; system sizes $N = 1000, 2000$, and 4000. We varied the sparsity parameter $\varepsilon \in \{.01, .05, .10, .15, .20, .25\}$ and, at each fixed sparsity level we varied the undersampling parameter δ around the predicted phase transition level, according to $M(\varepsilon|\eta) + \{-.015, -.010, -.005, 0.00, .005, .010, .015\}$ with $M(\varepsilon|\eta)$ being the minimax MSE; thus we are choosing parameters in the region where the formula (1.6) predicts a transition.

We defined success according to whether or not

$$\frac{\|x^0 - x^{300}\|_2^2}{\|x^0\|_2^2} < 0.01.$$

- *Monotone sequences.* For monotone regression, we used the pool-adjacent-violators algorithm. For determining the AMP-mono phase transition, we ran Monte Carlo simulations with the following parameters: 1000 Monte Carlo repetitions at each parameter combination; system sizes $N = 200$ (only, because of relative computational slowness). We varied the sparsity parameter $\varepsilon \in \{.01, .05, .10, .15, .20, .25\}$ and, at each fixed sparsity level we varied the undersampling parameter δ around the predicted phase transition level, according to $M(\varepsilon|mono) + \{-.015, -.010, -.005, 0.00, .005, .010, .015\}$ with $M(\varepsilon|mono)$ being the worst-case MSE as discussed in Appendix E.

We found that it was necessary to define success differently in this experiment than it had been defined for separable nonlinearities. We found that the relative *MSE* did not exhibit a sharp ‘first order’ phase transition, but rather seemingly a second order transition, which seems hard to locate precisely. We found that the success measure

$$N^{-1} \#\{i : |x_0(i) - x^{300}(i)| > 0.01\} < 0.01$$

gave relatively sharp phase transitions which could be located precisely.

To obtain the phase transition for the convex optimization problem (P_{mono}) we solved (P_{mono}) using CVX.

- *TV Minimization.* For TV minimization, we used the software package `tvdpip`. We considered many other software packages but found that they were not adaptable to the $1 - d$ case or else not reliable. We ran Monte Carlo simulations with the following parameters: 50 Monte Carlo repetitions at each parameter combination; system sizes $N = 200$ (only, because of relative computational slowness). We varied the sparsity parameter $\varepsilon \in \{.01, .05, .10, .15, .20, .25\}$

γ	R^2
1/3	0.99521
1/2	0.98724
2/3	0.9666811
3/4	0.950506
1	0.893454

Table 4: Powers γ and resulting R^2 for fitting Model (H.1)

and, at each fixed sparsity level we varied the undersampling parameter δ around the predicted phase transition level, according to $M(\varepsilon|TV) + \{-0.015, -0.010, -0.005, 0.00, .005, .010, .015, .020, .025\}$ with $M(\varepsilon|TV)$ being the worst-case MSE as discussed in Appendix E.

As in the monotone regression case, it was necessary to define success differently here than for separable nonlinearities. We again used the success measure

$$N^{-1} \#\{i : |x_0(i) - x^{300}(i)| > 0.01\} < 0.01.$$

To obtain the phase transition for the convex optimization problem (P_{tv}) we used CVX.

H Finite- N Scaling

The empirical phase transitions observed in the scalar separable case admit further analysis, to study whether (a) the offsets tend towards zero with increasing N (as we predict) and whether (b) the steepness of the phase transition increases with increasing N .

H.1 Offsets Decay Toward Zero

As described in the text, at each fixed value ε of the sparsity parameter, we gathered data at several different values of δ , and obtained the empirical phase transition parameter $\widehat{\text{PT}}(N, \varepsilon, \eta)$ (say), recorded in units of *offset* from prediction, so that $\widehat{\text{PT}}(N, \varepsilon, \eta) = 0$ means that the 50% success location fitted to the ε -fixed, δ -varying dataset is exactly at the predicted location $M(\varepsilon|\eta)$. Our analysis gave not only the empirical phase transition location, but also its formal standard error $SE(\widehat{\text{PT}})$.

We fit a linear model to the dataset, considering a range of powers $N^{-\gamma}$ that might be describing the decay of the offset with increasing N :

$$\widehat{\text{PT}}(N, \varepsilon, \eta) = c(\varepsilon, \eta)N^{-\gamma} + SE(\widehat{\text{PT}}) \cdot \text{N}(0, 1), \quad (\text{H.1})$$

Table 4 shows that $\gamma = 1/3$ provides an adequate description of the offsets, with an R^2 exceeding 0.995. A plot of raw $\widehat{\text{PT}}$'s versus the predictions of model (H.1) is given in Figure 23.

H.2 Transitions Sharpen

In addition to an empirical phase transition parameter $\widehat{\text{PT}}(N, \varepsilon, \eta)$ we also fitted an empirical steepness parameter $\hat{\beta}(N, \varepsilon)$, according to the logistic model:

$$\text{logit}(\text{SuccessProb}(N, \varepsilon, \delta, \eta)) = \hat{\beta}(N, \varepsilon, \eta)(\delta - \widehat{\text{PT}}(N, \varepsilon))$$

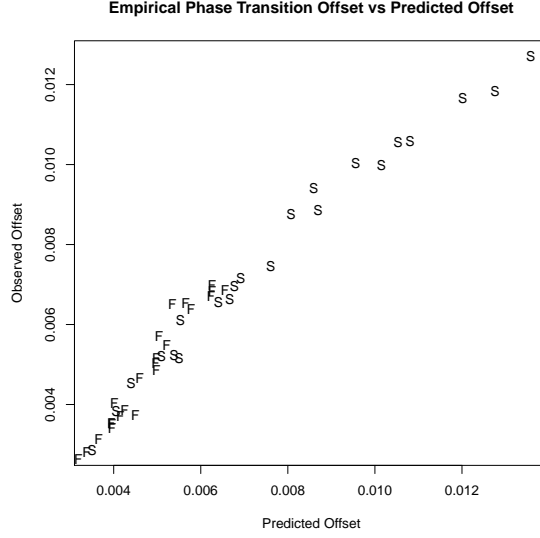


Figure 23: Fitted Offsets versus Predicted Offsets in model (H.1)

γ	R^2
1/3	0.99286
1/2	0.99927
2/3	0.991064
3/4	0.982350
1	0.947641

Table 5: Powers γ and resulting R^2 for fitting Model (H.2)

Here we expect $\hat{\beta}$ to grow with increasing N , corresponding to increasingly abrupt transitions from complete failure at $\delta \ll \hat{PT}(N, \varepsilon)$ to complete success at $\delta \gg \hat{PT}(N, \varepsilon)$. We fit a linear model to the $\hat{\beta}$ data considering a range of powers $N^{+\gamma}$ that might be describing the growth of the steepness with increasing N :

$$\hat{\beta}(N, \varepsilon, \eta) = c(\varepsilon, \eta)N^\gamma + \text{Error}, \quad (\text{H.2})$$

(We did not do a weighted least squares fit here. Although we should have done so, it will be evident that the results probably wouldn't much change had we done so).

Table 5 shows that $\gamma = 1/2$ provides an adequate description of the steepnesses, with an R^2 exceeding 0.999. A plot of raw $\hat{\beta}$'s versus the predictions of model (H.2) is given in Figure 24.

I Risk Function Properties

In this appendix we prove several useful properties of the block soft and James-Stein denoisers.

[When $d = 1$, this reduces to the derivative formula for scalar thresholding.] For the upper bound (I.2), use the Poisson mixture representation $f_{\xi,d+2}(w) = \sum_0^\infty p_{\xi/2} f_{d+2+2j}(w)$ and an identity for the central χ^2 density family to obtain

$$\frac{f_{d+2+2j}(w)}{w} = \frac{f_{d+2j}(w)}{d+2j} \leq \frac{f_{d+2j}(w)}{d}$$

and so to conclude that $f_{\xi,d+2}(w) \leq (w/d)f_{\xi,d}(w)$.

Hence the second term in (I.5) is bounded by $(1 - 1/\tau) \int_{\tau^2}^\infty f_{\xi,d}(w) dw$, whence follows the conclusion $(\partial r / \partial \xi) \leq 1$. Property (I.3) is immediate from (I.1) and (I.2) and the large- ξ limit of r .

The risk function at zero can be evaluated exactly via

$$R(0; \tau) = \int_{\tau^2}^\infty (\sqrt{w} - \tau)^2 f_d(w) dw = \sum_{k=0}^2 c_k \tilde{F}_{d+k}(\tau^2)$$

for $c_0 = \tau^2$, $c_1 = -2\sqrt{2}\Gamma((d+1)/2)/\Gamma(d/2)$ and $c_2 = d$. To obtain the bound in (I.4), write the risk function using the unbiased risk formula as

$$R(\mu) = \xi + \mathbb{E}[2d + \tau^2 - S^2 - 2(d-1)\tau S^{-1}, S > \tau]$$

and observe that the integrand is decreasing in S , and bounded above by 2 for $S > \tau$, so that $R(0) \leq 2\mathbb{P}(\chi_d^2 \geq \tau^2)$. This yields (I.4). □

I.2 Positive-Part James-Stein Shrinkage

Let e_1 denote the vector $(1, 0, \dots, 0)$ in \mathbf{R}^d and $Y \sim \mathbf{N}(\mu \cdot e_1, I_{d \times d})$ and put $S = \|Y\|_2^2$. We have the noncentral chi-squared distribution $S \sim \chi_d^2(\xi)$ with noncentrality $\xi = \mu^2$.

Let $\eta^{JS}(x) = (1 - (d-2)/S)_+ \cdot Y$. Stein's unbiased estimate of risk is

$$U(S) = \begin{cases} S - d & S < d - 2, \\ d - (d-2)^2/S & S > d - 2. \end{cases}$$

Hence,

$$R(\mu) = \mathbb{E}U(S).$$

I.2.1 Risk at 0

For central chi-squared on d degrees of freedom, the density satisfies $wf_d(w) = df_{d+2}(w)$. With $\tilde{F}_d(x) = P(\chi_n^2 \geq x)$,

$$\begin{aligned} R(0) &= \mathbb{E} \left\{ \left\| \left(1 - \frac{d-2}{S}\right)_+ Y \right\|^2 \right\} \\ &= \int_{d-2}^\infty \left[w - 2(d-2) - \frac{(d-2)^2}{w} \right] f_d(w) dw \\ &= d\tilde{F}_{d+2}(d-2) - 2(d-2)\tilde{F}_d(d-2) + (d-2)\tilde{F}_{d-2}(d-2). \end{aligned}$$

With $D = d - 2$, we can rewrite the second line as

$$R(0) = D^{-1} \mathbb{E}(\chi_D^2 - D)_+^2.$$

We have the convergence in distribution,

$$D^{-1/2}(\chi_D^2 - D) \Rightarrow \mathbf{N}(0, 1), \quad D \rightarrow \infty,$$

adding a tightness argument (not given here) one can show that

$$R(0; d) \rightarrow 2\mathbb{E}Z_+^2 = 1, \quad d \rightarrow \infty,$$

where $Z \sim \mathbf{N}(0, 1)$. Numerically $R(0; d) \approx 1 + 0.752/\sqrt{d}$ for large d .

I.2.2 Monotonicity of $R(\mu)$.

Here and in the next subsection we use the variation-diminishing version of total positivity:

Theorem I.1. *The non-central χ^2 family is strictly variation diminishing of all orders*

See [BJM81].

Let $S^-(g)$ and $S^+(g)$ denote the number of sign changes and strict sign changes of function g on $[0, \infty)$, and let $IS^-(g)$ and $IS^+(g)$ denote the initial sign (at 0). Given a function $g : \mathbf{R}_+ \mapsto \mathbf{R}_+$, define function $\gamma : \mathbf{R}_+ \mapsto \mathbf{R}_+$ by

$$\gamma(\xi) = \int g(w) f_{d,\xi}(w) dw.$$

The SVR property says that $S^+(\gamma) \leq S^-(g)$ and that if $S^+(\gamma) = S^-(g)$ then necessarily $IS^+(\gamma) = IS^-(g)$.

Now verify that the risk $R(\mu)$ of η^{JS} is monotone increasing in $\mu > 0$. Put $G(\xi) = \mathbb{E}_{\mu e_1} U(S)$. Since $s \rightarrow U(s)$ is strictly increasing, it follows from SVR₂ that so also is $G(\xi)$. Indeed, consider $g = U - a$ for all constants a , and infer that the number of sign changes in the corresponding γ is not larger than the number in g . But $\gamma = G - a$, and so G is monotone. Finally: $G(\xi) = R(\mu)$ where $\xi = \mu^2$.

I.2.3 Superquadraticity of $R(\mu)$.

Lemma I.2. *For each $a > 0$, the function $\mu^2 \rightarrow R(\mu) - a\mu^2$ has one strict sign change.*

This implies superquadraticity: given μ_0 , set $a = R(\mu_0)/\mu_0^2$. Obviously $R(\mu) - a\mu^2$ changes sign at μ_0 , but the lemma says this is the only change. At the same time, we know $R(0) > 0$ by a previous subsection. Hence, $R(\mu) - a\mu^2 > 0$ for $\mu < \mu_0$, and so

$$R(\mu) > \left(\frac{\mu}{\mu_0}\right)^2 R(\mu_0), \quad 0 \leq \mu < \mu_0.$$

Proof. Since $\mathbb{E}[S - d] = \xi = \mu^2$, we have

$$\gamma_a(\xi) = R(\mu) - a\mu^2 = \mathbb{E}_\xi g(S),$$

with

$$g(S) = \begin{cases} (1-a)(S-d) & S < d-2, \\ d - (d-2)^2 S^{-1} - a(S-d) & S > d-2. \end{cases} \quad (\text{I.6})$$

By inspection of the function g in (I.6), we have

$$S^-(g) = \begin{cases} 2 & a < 1, \\ 1 & a \geq 1, \end{cases} \quad \begin{matrix} IS^-(g) = - \\ IS^-(g) = +. \end{matrix}$$

Note that $\gamma(0) = R(0) > 0$, so $IS^+(\gamma) = +$. Consequently $S^+(\gamma) \leq 1$ in both cases. Then from behavior at $+\infty$ it follows that $S^+(\gamma) = 1$. \square

References

- [BC83] P. Bickel and J. Collins, *Minimizing fisher information over mixtures of distributions*, Sankhya A **45** (1983), 1–19.
- [BCDH10] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, *Model-Based Compressive Sensing*, IEEE Trans. on Inform. Theory **56** (2010), 1982–2001.
- [BCT11] J. D. Blanchard, C. Cartis, and J. Tanner, *Compressed sensing: How sharp is the restricted isometry property?*, SIAM Review **53** (2011), no. 1, 105–125.
- [BGI⁺08] R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss, *Combining geometry and combinatorics: A unified approach to sparse signal recovery*, 47th Annual Allerton Conference (Monticello, IL), September 2008, pp. 798 – 805.
- [Bic81] P. Bickel, *Minimax estimation of the mean of a normal distribution when the parameter space is restricted*, Annals of Statistics **9** (1981), 1301–1309.
- [BJM81] L.D. Brown., I.M. Johnstone, and K.B. MacGibbon, *Variation diminishing transformations: A direct approach to total positivity and its statistical applications*, Journal of the American Statistical Association **76** (1981), no. 376, 824–832.
- [BLM12] M. Bayati, M. Lelarge, and A. Montanari, *Universality in message passing algorithms*, In preparation, 2012.
- [BM11a] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. on Inform. Theory **57** (2011), 764–785.
- [BM11b] ———, *The LASSO risk for gaussian matrices*, IEEE Trans. on Inform. Theory (2011), arXiv:1008.2581.
- [BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Amer. J. of Mathematics **37** (2009), 1705–1732.

- [BT10] J. Blanchard and J. Tanner, *Large scale iterative hard thresholding on a graphical processing unit*, ICNAAM 2010: International Conference of Numerical Analysis and Applied Mathematics **1281** (2010), 1730–1734.
- [Cai99] T.T. Cai, *Adaptive wavelet estimation: A block thresholding and oracle inequality approach*, The Annals of Statistics **27** (1999), no. 3, pp. 898–924.
- [Can06] E. Candès, *Compressive sampling*, Proceedings of the International Congress of Mathematicians (Madrid, Spain), 2006.
- [CHDB08] V. Cevher, C. Hegde, M. F. Duarte, and R. G. Baraniuk, *Sparse Signal Recovery Using Markov Random Fields*, Neural Information Processing Systems (Vancouver), December 2008.
- [CICB10] V. Cevher, P. Indyk, L. Carin, and R.G. Baraniuk, *Sparse Signal Recovery and Acquisition with Graphical Models*, IEEE Signal Processing Magazine **27** (2010), 92–103.
- [CRPW11] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A.S. Willsky, *The convex geometry of linear inverse problems*, arXiv:1012.0621, 2011.
- [CS81] G. Casella and W. E. Strawderman, *Estimating a bounded normal mean*, Annals of Statistics **9** (1981), 870–878.
- [CT05] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. on Inform. Theory **51** (2005), 4203–4215.
- [CY08] R. Chartrand and W. Yin, *Iteratively reweighted algorithms for compressive sensing*, ICASSP, IEEE, 2008, pp. 3869–3872.
- [DDS92] J.C. Hoch D.L. Donoho, I.M. Johnstone and A.S. Stern, *Maximum entropy and the nearly black object (with discussion)*, J. Roy. Statist. Soc. B **54** (1992), 41–81.
- [DJ94] D. L. Donoho and I. M. Johnstone, *Minimax risk over l_p balls*, Prob. Th. and Rel. Fields **99** (1994), 277–303.
- [DJ95] D.L. Donoho and I.M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, J. Amer. Statist. Assn. **90** (1995), 1200–1224.
- [DMM09] D. L. Donoho, A. Maleki, and A. Montanari, *Message Passing Algorithms for Compressed Sensing*, Proceedings of the National Academy of Sciences **106** (2009), 18914–18919.
- [DMM10] ———, *Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction*, Proceedings of IEEE Inform. Theory Workshop (Cairo), 2010.
- [DMM11] D.L. Donoho, A. Maleki, and A. Montanari, *The Noise Sensitivity Phase Transition in Compressed Sensing*, IEEE Trans. on Inform. Theory **57** (2011), 6920–6941.
- [Don06] D. L. Donoho, *High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension*, Discrete Comput. Geometry **35** (2006), 617–652.

- [DT05] D. L. Donoho and J. Tanner, *Neighborliness of randomly-projected simplices in high dimensions*, Proceedings of the National Academy of Sciences **102** (2005), no. 27, 9452–9457.
- [DT09a] ———, *Counting faces of randomly projected polytopes when the projection radically lowers dimension*, Journal of American Mathematical Society **22** (2009), 1–53.
- [DT09b] D.L. Donoho and J. Tanner, *Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing*, Phil. Trans. R. Soc. A **367** (2009), 4273–4293.
- [DT10a] D. L. Donoho and J. Tanner, *Counting the faces of randomly-projected hypercubes and orthants, with applications*, Discrete & Computational Geometry **43** (2010), no. 3, 522–541.
- [DT10b] D. L. Donoho and J. Tanner, *Counting the faces of randomly-projected hypercubes and orthants, with applications*, Discrete & Computational Geometry **43** (2010), no. 3, 522–541.
- [DT10c] D. L. Donoho and J. Tanner, *Precise undersampling theorems*, Proceedings of the IEEE **98** (2010), 913–924.
- [GB97] H.-Y. Gao and A.G. Bruce, *Waveshrink with firm shrinkage*, Statistica Sinica **7** (1997), 855–874.
- [HKP98] P. Hall, G. Kerkycharian, and D. Picard, *Block threshold rules for curve estimation using kernel and wavelet methods*, Ann. Statist. **26** (1998), 922–942.
- [HR09] P.J. Huber and E. Ronchetti, *Robust statistics (second edition)*, J. Wiley and Sons, 2009.
- [Hub64] P.J. Huber, *Robust estimation of a location parameter*, The Annals of Mathematical Statistics **35** (1964), no. 1, 73–101.
- [Joh94a] I. M. Johnstone, *Minimax bayes, asymptotic minimax and sparse wavelet priors*, Statistical Decision Theory and Related Topics, V (S.S. Gupta and eds. J.O. Berger, eds.), 1994, p. 303326.
- [Joh94b] I.M. Johnstone, *On minimax estimation of a sparse normal mean vector*, Annals of Statistics **22** (1994), 271289.
- [JS10] W. James and C. Stein, *Estimation with quadratic loss*, Proc. Fourth Berkeley Symp. Math. Statist. Prob (Berkeley), 2010.
- [KGR11] U.S. Kamilov, V.K. Goyal, and S. Rangan, *Message-Passing Estimation from Quantized Samples*, arXiv:1105.6368, 2011.
- [KWT09] Y. Kabashima, T. Wadayama, and T. Tanaka, *A typical reconstruction limit for compressed sensing based on l_p -norm minimization*, J.Stat. Mech. (2009), L09003.

- [KXAH10] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi, *Improved sparse recovery thresholds with two-step reweighted ℓ_1 minimization*, arXiv:1004.0402.
- [Mal78] C. L. Mallows, *Minimizing an integral*, SIAM Review **20** (1978), 183.
- [MAYB11] A. Maleki, L. Anitori, A. Yang, and R. Baraniuk, *Asymptotic Analysis of Complex LASSO via Complex Approximate Message Passing (CAMP)*, arXiv:1108.0477, 2011.
- [MD10] A. Maleki and D. L. Donoho, *Optimally tuned iterative reconstruction algorithms for compressed sensing*, IEEE Journal of Selected Topics in Signal Processing **4** (2010), 330–341.
- [MM09] M. Mézard and A. Montanari, *Information, Physics and Computation*, Oxford University Press, Oxford, 2009.
- [Mon12] A. Montanari, *Graphical models concepts in compressed sensing*, Compressed Sensing (Y.C. Eldar and G. Kutyniok, eds.), Cambridge University Press, 2012.
- [Ran11] S. Rangan, *Generalized Approximate Message Passing for Estimation with Random Linear Mixing*, IEEE Intl. Symp. on Inform. Theory (St. Perersbourg), August 2011.
- [ROF92] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena **60** (1992), 259–268.
- [RRN11] N. Rao, B. Recht, and R. Nowak, *Tight measurement bounds for exact recovery of structured sparse signals*, arXiv:1106.4355, 2011.
- [RU08] T.J. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, Cambridge, 2008.
- [Sch10] P. Schniter, *Turbo Reconstruction of Structured Sparse Signals*, Proceedings of the Conference on Information Sciences and Systems (Princeton), 2010.
- [Sch11] ———, *A message-passing receiver for BICM-OFDM over unknown clustered-sparse channels*, arXiv:1101.4724, 2011.
- [SPS10] S. Som, L. C. Potter, and P. Schniter, *Compressive Imaging using Approximate Message Passing and a Markov-Tree Prior*, Proc. Asilomar Conf. on Signals, Systems, and Computers, November 2010.
- [SSS10] L.C. Potter S. Som and P. Schniter, *On Approximate Message Passing for Reconstruction of Non-Uniformly Sparse Signals*, Proceedings of the National Aereospace and Electronics Conference (Dayton, OH), 2010.
- [Sto09] M. Stojnic, *Block-length dependent thresholds in block-sparse compressed sensing*, arXiv:0907.3679, 2009.
- [TSR⁺05] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, J. R. Statist. Soc. B **67** (205), 91–108.

- [VO96] C. R. Vogel and M. E. Oman, *Iterative methods for total variation denoising*, SIAM J. Sci. Comput **17** (1996), 227–238.
- [Wai09] M.J. Wainwright, *Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso)*, IEEE Trans. on Inform. Theory **55** (2009), 2183–2202.
- [XH10] W. Xu and B. Hassibi, *Compressive sensing over the grassmann manifold: a unified geometric framework*, arXiv:1005.3729, 2010.
- [YL10] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society B **68** (2010), 49–67.